성문특성이 제거된 성도특성 추출에 관한 연구

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

A Study on Extract of Vocal Tract Characteristic after Concealing the Vocal Cord Property

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

요 약

Since the amplitude of voiced fall off at about -20dB/decade, dynamic range is often compressed prior to spectral analysis so that details at weak, high frequencies may be visible. Preemphasizing the speech, either by differentiating the analog speech $s_a(t)$ prior to A/D conversion or by differencing the discrete-time $s(n) = s_a(nT)$, compensating for falloff at high frequencies. The most common form of preemphasis is $y(n) = s(n) - As(n-1)$, where A typically lies between 0.9 and 1.0 and reflects the degree of pre-emphasis. In this paper, we proposed that A is adjusted at each time by measuring the slope of envelope in frequency domain.

1. 서 론

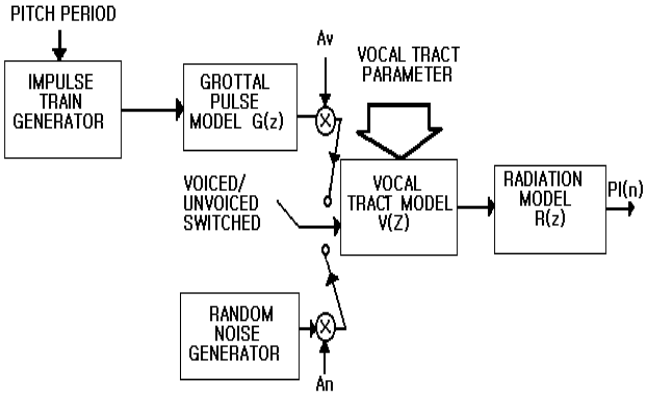
음성신호를 관찰했을 때 성문특성으로 인해 고주파쪽 특성이약화되는경향이있다.이를보상하기위해 Preemphasis 필터가 사용되어진다. Preemphasis 필터를 간단히 수식으로 표현하면 $y(n) = s(n) - As(n-1)$와 같이 차분방정식으로 나타낼 수가 있다[5][6]. 여기서 A값은 보통 0.9에서 1 사이의 값을 주로 사용한다. 본 논문에서는 주파수 영역에서 포락선 기울기를 측정하여 이에 따라 A값을 각 프레임별로 특성에 맞게 보상해줄 것을 제안한다. 2장에서는 Preemphasis 필터와 음성생성 모델에 대하여 설명하고 3장에서는 본논문에서 주파수 영역의 포락선 기울기 측정에 사용한 방법을 간단한 수식을 통해서 살펴본다. 4장에서는 실험 및 결과, 5장에서는 결론을 맺는다.

2. 음성신호의 생성모델

2-1. 음성신호의 생성

음성신호는 소리와 반복으로 이루어진다. 소리와 그 사이의 변이는 정보에 대한 기호적인 표현으로 나타난다. 소리에 대한 배열은 언어의 규칙에 의해 결정되며, 이 규칙과 인간의 통신에 있어서 의미에 대한 연구는 언어학의 영역이며, 음성의 소리를 연구하고 분류하는 것은 음성학 분야이다[1][2].

음성신호의 구조에 대한 연구는 음성정보를 추출하거나 강조할 수가 있다. 따라서 음성신호의 생성에 대한 수학적 모델은 음성을 처리하는데 있어서 매우 중요한 영역이다. 성도(vocal tract)는 성대(vocal cord)와 입술 끝까지를 말한다. 따라서 성도는 인두(식도에서 입을 연결하는 부위)와 입 또는 구강으로 구성된다. 남성의 성도 길이는 평균 17㎝ 정도이다. 성도의 단면적은 혀, 입술, 턱 그리고, 0㎠(완전히 닫혔을 때)에서 약 20㎠까지 변화하는 연구개의 위치에 의해 결정된다. 비도(nasal tract)는 연구개에서 시작하여 콧구멍에서 끝난다. 연구개가 낮아질 때 비도는 비음을 생성하기 위해 음향학적으로 성도에 연결된다. 음성은 간단히 공기가 허파로부터 방출되고 결과적으로 성도에 있는 협착점에 의해 공기가 동요될 때 이 시스템으로부터 방사되는 음향학적 파형이다[1][2]. 음성은 여기(excitation)에 따라 세 가지로 나눌 수 있다. 첫째, 유성음은 조정된 성대의 팽창과 함께 성문을 통한 공기의 힘에 의해 생성되어 감쇄 진동하여 성도를 자극하는 공기의 준 주기적인 펄스를 만든다. 유성음은 /아/, /에/, /이/, /오/, /우/ 등의 모음과 /ㄹ/, /ㅁ/등의 비음으로 구성된다. 둘째, 마찰음 또는 무성음은 성도에 있는 어떤 점에서 협착을 형성하고, 동요를 만들기 위해 고속으로 협착점을 통과하는 공기의 힘에 의해 발생된다. 이것은 성도를 자극하기 위해 광대역 잡음원을 생성하게 된다. 셋째, 파열음(plosive)은 완전히 입을 폐쇄하고, 이 폐쇄 뒤에서 압력을



[그림 1] 일반적인 음성생성에 대한 이산시간 모델

만들어 갑자기 느슨하게 함으로써 생성된다. 음성생성에 있어서 성도의 공명주파수를 포만트 주파수 또는 포만트라고 한다. 포만트 주파수는 성도의 모양과 면적에 따라 다르고, 소리의 형태는 성도의 모양이 변화함으로써 시간에 따라 변한다.

2-2. 일반적인 음성생성모델

음성생성에 대한 선형모델은 50년대 후반 Fant에 의해 개발되었는데 그는 음성출력을 음원이 여파기를 통과하여 나오는 신호로 가정하고, 음원과 성도의 각 부분을 독립적인 것으로 간주하는 선형예측모델을 제시하였다. 음원에 대한 모델로 유성음의 음원은 준주기적인 펄스, 무성음의 음원은 백색잡음을 사용하였고, 성대에서 성문이 음원에 미치는 영향은 다음과 같은 성문모델(GlottalShaping model)로 모델링하였다[1][2].

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2} \qquad (1)$$

여기서, T는 준주기이고, cT는 감소인자이며 1 보다 충분히 작다.

성문을 지난 신호는 성도를 거치면서 성도의 형태에 따라 몇 개의 공명주파수를 갖게 되는데 이 공명주파수와 대역폭을 2-극(pole) 여파기로 나타내면 성도에 대한 모델은 다음과 같이 구할 수 있다.

$$V(z) = \frac{1}{\Pi^{k}[1 - 2e^{-ciT}\cos(BiT) z^{-1} + e^{-2ciT} z^{-2}]} \qquad (2)$$

여기서, k는 포만트의 개수이고, c_i와 B_i는 포만트 주파수와 대역폭을 결정하는 값으로 대역폭의 좁다는 가정에서 실제 포만트의 공명주파수와 대역폭은 각각 $Bi/2\pi$, $ci/2\pi$로 결정된다.

성도를 통과한 신호는 마지막으로 입술을 통과하는데 입술에서의 방사특성에 대한 모델은 다음과 같다.

$$L(z) = 1 - z^{-1} \qquad (3)$$

음원에 대한 모델을 E(z)로 하면 음성출력 X(z)는 다음과 같다.

$$X(z) = E(z)G(z)V(z)L(z) \qquad (4)$$

이를 전극(all-pole)형 합성모델로 다시 정의하면,

$$X(z) = E(z)\frac{1}{A(z)} \qquad (5)$$

여기서,

$$A(z) = \frac{1}{G(z)V(z)L(z)} \qquad (6)$$

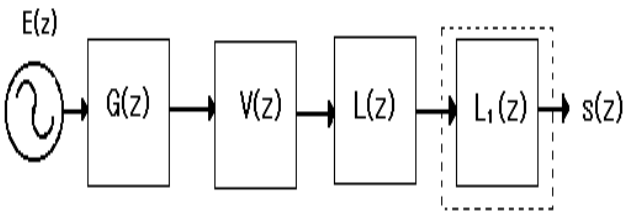
이 A(z)는 다음 관계식에 의해 음성으로부터 성도특성을 제거하고 음원을 이끌어 내는 가역여파기임을 알 수 있다.

$$E(z) = X(z)A(z) \quad (7)$$

이상에서 살펴본 것과 같이 음성을 음원과 그 음원이 통과하는 성도여파기로 모델링하고 각각을 독립적으로 모델링함으로써 수학적으로 선형방정식의 해를 구할 수 있다[1][2].

2-3. 성문특성을 고려한 음성생성모델

2-2절에서는 일반적인 음성생성모델에 대하여 알아보았다. 위에서 설명했듯이 입술에서의 방사특성은 1 zero로 표현 가능하다. 그림 2에서는 성문특성을 고려한 음성생성모델을 나타내었다.



[그림 2] 본 논문에서 고려한 음성생성모델

그림 2에서 E(z)은 여기신호, G(z)은 2 pole로 모델링이 가능한 성문특성, V(z)은 성도특성, 즉 주파수 영역에서의 포만트 특성을 나타낸다. 그리고 L(z)은 입술의 방사특성을 나타내는데 마이크로폰을 가깝게 대고 발음을 한 경우에 있어서는 커플링(Coupling)이 발생하여 Zero 특성이 없어지게 된다. 마이크로폰을 멀리대고 발음한 순간에는 그림 2에서 성문특성을 나타내는 2 pole중 하나가 1 zero의 L(z)와 상쇄되어 없어지므로 성문특성에 의해 영향을 받

는 성도특성을 제대로 관찰하기 위해서는 그림 2에서 처럼 $L_1(z)$을 사용하여 보상을 시켜주어야 한다. 위에서 설명한 내용을 간단히 수식을 통해 살펴보면 다음과 같다.

$$S(z) = E(z) \cdot G(z) \cdot V(z) \cdot L(z) \quad (8)$$

식 (8)은 마이크로폰을 멀리대고 발성을 한경우를 나타낸다. 그러므로 2 pole로 모델링한 성문특성이 1 zero로 모델링 되는 입술의 방사특성으로 인해 전체적으로 1 pole로 나타낼수가 있고, 대략적으로 주파수 영역에서의 기울기는 -20dB/decade 정도로 나타난다[1][3][4].

$$S'(z) = L_1(z) \cdot S(z) = E(z) \cdot V(z) \quad (9)$$

식 (9)는 성도특성의 정확한 분석을 위해 성도특성에 미치는 성문특성을 제거하기 위한 경우를 나타낸 식이다. 식 (9)는 다음과 같이 표현 가능하다.

$$V(z) \cong \frac{S'(z)}{E(z)} \quad (10)$$

즉, 식 (8)~식 (10)을 통해서 마이크로폰의 위치에 따라 주파수 영역의 포만트특성의 기울기가 변화함을 알수가 있다.

2-4. PRE-EMPHASIS 필터의 특징

음성신호처리 분야에서 스펙트럼 경사(Spectrum Tilt)를 평탄화해 줌으로써 신호의 동적범위(Dynamic Range)를 억제하는 프리엠퍼시스 과정은 SNR을 높이는데 유효한 것으로 알려져 있다[5][6]. 이 방법은 일반적으로 A/D 변환을 위한 저역통과 필터링에 앞어서 20dB/decade 정도의 고주파 영역을 강조하는 역할을 한다. 이 방법은 또한 A/D 변환 다음에도 수행되어질 수 있는데, 차분방정식이나 식 (11)과 같은 1차 디지털 필터링을 통하여 구현된다.

$$H(z) = 1 - az^{-1} \quad (11)$$

여기에서 a값은 대략 1에 가까운 값을 갖는다. SNR을 가능한 높이기 위해서는 A/D 변환을 하기 앞서 프리엠퍼시스를 해야한다. 그리고 원신호의 스펙트럼 경사를 복구하기 위한 과정을 디엠퍼시스(deemphasis)라고 한다.

3. 주파수영역에서의 기울기 측정

단구간 자기상관 함수는 식 (12)로 표현 가능하다[8].

$$\Phi_n(i,j) = \sum_{m=0}^{N-1-(i-j)} s_n(m) s_n(m+i-j),\ 1 \le i \le p,\ 0 \le j \le p \quad (12)$$

여기서 $R_n(j) = \sum_{m=0}^{N-1-j} s_n(m) s_n(m+j)$ (13)

$$\sum_{j=1}^{p} a_j \Phi_n(i,j) = \Phi_n(i,0),\ \text{for}\ i = 1, \ldots, p \quad (14)$$

자기상관법(Auto-correlation Method)을 이용하여 식 (14)를 풀면 다음과 같이 표현된다.

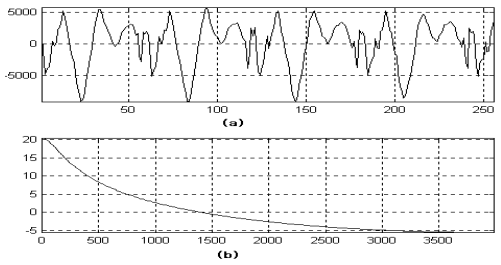
$$\begin{vmatrix} R_n(0) & R_n(1) & \cdot & R_n(p-1) \\ R_n(1) & \cdot & \cdot & R_n(p-2) \\ \cdot & \cdot & \cdot & \cdot \\ R_n(p-1) & \cdot & \cdot & R_n(0) \end{vmatrix} \begin{vmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \alpha_p \end{vmatrix} = \begin{vmatrix} R_n(1) \\ R_n(2) \\ \cdot \\ R_n(p) \end{vmatrix} \quad (15)$$

p=1에 대하여 위의 식을 정리하면 다음과 같은 식으로 표현 가능하다[8].

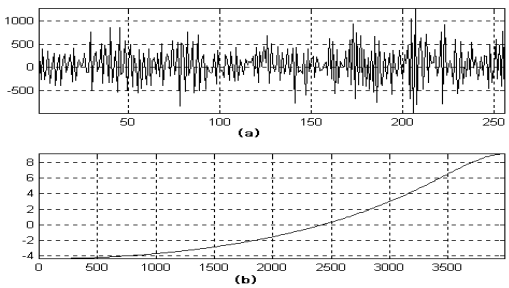
$$\alpha_1 = \frac{R_n(1)}{R_n(0)} \quad (16)$$

4. 실험 및 결과

실험을 위해 IBM PC(333 MHz)에 마이크 입력이 가능한 A/D 변환기를 인터페이스 하였다. 음성시료는 일기예보 뉴스에서 발췌한 남/녀 아나운서의 음성을 8kHz로 표본화하고 16bit로 양자화하여 사용하였다.
그림 3, 그림4는 음성신호와 상호상관 함수를 이용하여 측정한 음성신호의 기울기를 나타낸것이다.



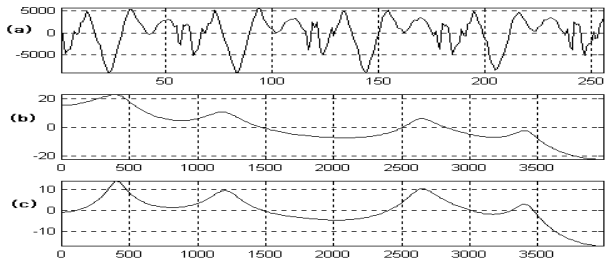
[그림 3] 음성신호(a), 측정된 기울기(b)



[그림 4] 음성신호(a), 측정된 기울기(b)

그림 5와 그림 6은 주파수 영역에서 프레임 단위로 기울기를 측정하여 프리엠퍼시스 필터의 A값의 조정을 통하여 나타난 주파수 영역의 스펙트럼이다. 그림 7과 그림

8은 각 프레임별 기울기를 측정하여 이에따라 프리엠퍼시스의 A값을 조절한 LPC이다.



[그림 5] 기울기에 측정하여 A값 조절을 통한 결과(1)
(a)음성신호 (b) 스펙트럼분석
(c)기울기에 따라 A값을 조절한 스펙트럼분석

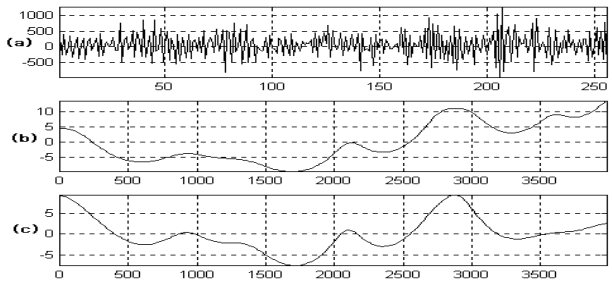
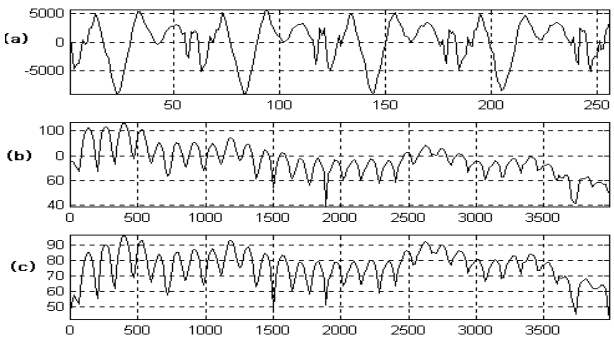
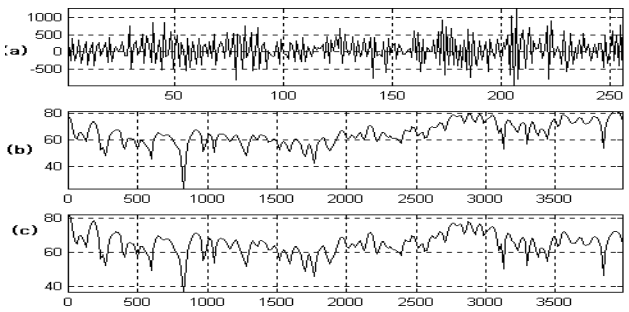


그림 6. 기울기에 측정하여 A값 조절을 통한 결과(2)



[그림 7] LPC 분석(1)
(a)음성신호 (b)LPC분석 (c)평탄화된 결과



[그림 8] LPC 분석(2)
(a)음성신호 (b)LPC분석 (c)평탄화된 결과

5. 결 론

음성신호를 관찰했을때 성문특성으로 인해 고주파 특성이 약화되는 경향이 있다. 이를 보상하기 위해 Preemphasis필터가 사용되어진다. Preemphasis필터를 간단히 수식으로 표현하면 $y(n)=s(n)-As(n-1)$와 같이 차분방정식으로 나타낼수가 있다[5][6]. 여기서 A값은 보통 0.9에서 1사이의 값을 주로 사용한다. 본 논문에서는 자기상관함수를 이용하여 각 프레임별로 기울기를 측정하였다. 측정된 기울기를 이용하여 프레임 단위로 A값을 달리 조정하여 보상한 결과, 각각의 프레임 구간 특성에 따라 고주파 특성, 혹은 저주파 특성을 강조함을 알수가 있었다. 실험결과의 분석을 위해 LPC 분석을 한 결과, 좀더 첨예한 포만트를 관찰할 수가 있었는데 이는 기울기를 사용하여 A값을 조정하여 사용한 프리엠퍼시스 필터의 사용으로 성도특성을 제대로 tracking하고 있음을 반영한다.

참고문헌

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

[2] 배명진,"디지탈 음성분석", 동영출판사, 1998. 4.

[3] Oppenheim, Schafer,"Discrete Time Signal Processing", Prentice Hall, 1989.

[4] Emanuel C. Ifeachor,"Discrete Time Signal Processing", Addison Wesley,1993

[5] 오영환, "음성언어정보처리",홍릉과학출판사,1998.

[6] Douglas O, shaughnessy, "Speech Communication", IEEE Press,1996

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