1999년도 대한전기학회 정기총회 및 추계학술대회 논문집 1999. 11. 20

사이코어쿠스틱스 모델을 이용한 음성 향상

권철현, 신대규, 박상희
연세대학교 전기 · 컴퓨터 공학과

Speech enhancement using psychoacoustics model

Chul-Hyun Kwon, Dae-Kyu Shin, Sang-Hui Park
Dept. of Electrical&Computer Eng. Yonsei Univ.

Abstract - In this study, a speech enhancement is presented based on the utilization of well-known auditory mechanism, noise masking.
The speech enhancement approach adopted here is to derive an modifier that achieves audible noise suppression. This modification selectively affects the perceptually significant spectral values, and is therefore less prone to introduction of unwanted distortions than methods that affect the complete STSA and produces more enhanced results at low SNR as well as at high SNR.
The speech enhancement method adopted here needs exact estimation of the minimum specteal value per critical band because it uses only the minimum spectral value per critical band.
For this, the method adopted here uses the modified spectral subtraction that is more flexible than power spectral subtraction. So, the result in experiment represented better SNR than before.

1. 서 론

잡음으로 인해 손상된 음성을 향상시키는 문제는 지난 몇 십 년 동안 많은 방법들이 제시되어 왔지만 아직도 광범위하게 남아있다.

음성 향상을 위해 채택된 방법론에 따르면, 과거에는 음성의 주기성[1]-[2]이나 생성 매커니즘[3]과 같은 음성 신호 자체의 특수한 성질을 기본으로 하는 것들이었지만, 최근의 방법들은 주로 손상된 신호의 단시간 스펙트럴 크기(STSA)의 조작에 기본을 두고 있다.

본 논문에서 제시할 향상 방법은 잘 알려진 청각 메커니즘인 잡음 매스킹을 이용하는 것이다. 음성은 공존하는 잡음 성분들을 매스크할 수 있다. 이런 의미에서 듣는 사람이 인식하는 잡음 손상은 음성의 STSA의 시간에 따라 변하는 성질들에 따라 시계에서 변화할 것이다. 그것이 제거되어져야할 손상의 가청 잡음 성분이다.

본 논문에서 보일 향상을 위한 접근방법은 우선, 가청 잡음 제거가 가능한 모디파이어의 유도를 위해 확장되어 사용된 STSA의 가청 잡음 성분에 바탕을 둔다. 이러한 변형은 지각적으로 중요한 스펙트럴 값들에게 선택적으로 영향을 주어 전체적으로 STSA에 영향을 주는 방법들보다 더 강력하고 원하지 않는 왜곡들이 덜 나타나게 할 것이며, 무엇보다도 SNR이 높은 음성 신호에서 뿐만이 아니라 낮은 음성 신호에서도 높은 향상을 보이게 할 것이다[4].

이제, 음성 향상의 관점에서, 깨끗한 음성 신호에 대한 어떠한 추가 정보도 알려지지 않는다고 가정하면, 제안된 방법은 잡음이 첨가된 음성 신호로부터 크리티컬 밴드(critical band:CB)[5]당 스펙트럴 최소값의 정확한 추정에 달려있다. 이를 위해, 먼저 오래 전부터 사용되어 오는 파워 스펙트럴 서브트랙션(PSS)을 좀 더 유연한 형태의 변형된 스펙트럴 서브트랙션(MSS)[6]으로 일단 1차적인 음성 향상을 하였다. 그 결과 좀 더 나아진 SNR 향상이 가능해졌다.

2. 본 론

2.1 변형된 스펙트럴 서브트랙션

잡음 압축의 과정은 이득 함수 G(f)를 잡음이 첨가된 음성 신호 $|Y(f)|$에 곱함으로서 얻어진다.

$$|\hat{S}(f)| = G(f) \cdot |Y(f)| \qquad 0 \le G(f) \le 1 \tag{1}$$

만약에 $|\hat{D}(f)|^2 > |Y(f)|^2$이면, G(f)는 항상 실수값이기 때문에 G(f)는 0이 된다.

일반화된 스펙트럴 서브트랙션 다음과 같다[6].

$$G(f) = \begin{cases} \left(1-\alpha \cdot \left[\dfrac{|\hat{D}(f)|}{|Y(f)|}\right]^{\gamma_1}\right)^{\gamma_2}, \\ \quad \text{if } \left[\dfrac{|\hat{D}(f)|}{|Y(f)|}\right]^{\gamma_1} < \dfrac{1}{\alpha+\beta} \\ \left(\beta \cdot \left[\dfrac{|\hat{D}(f)|}{|Y(f)|}\right]^{\gamma_1}\right)^{\gamma_2} \end{cases} \quad (\gamma = \gamma_1 = 1/\gamma_2) \tag{2}$$

식(2)에서 파라미터들을 자유롭게 변화시켜 여러면에서 가장 좋은 결과를 가져오는 이득 함수를 구하면 된다.

2.2 인지적으로 중요한 스펙트라의 정의

추가된 잡음이 있을 경우, 잡음이 첨가된 음성 신호(y(n))는 원래의 깨끗한 음성 신호(x(n))와 잡음 성분(d(n))의 합으로 구성된다.

$$y(n) = x(n) + d(n), \quad 0 \le n \le N-1 \tag{3}$$

대부분의 실제적인 상황에서는 단시간 스펙트라가 필요하기 때문에 각각 $Y_w(k,i)$와 $X_w(k,i)$로 주어진 원도우된 y(n)와 x(n)의 FFT는 계산되어져야 한다. 각각에 대응하는 파워 스펙트라는 $Y_p(k,i)$와 $X_p(k,i)$로 주어진다.

사이코어쿠스틱 신호의 향상 기술의 기본 원리는 가청 잡음에 영향을 미치는 스펙트럴 성분을 압축하는 것이다. 이러한 성분들은 T(k,i)로 정의된 깨끗한 신호의 가청 매스킹 한계(audible masking threshold:AMT)의 추정으로부터 얻어질 수 있다.

잡음이 첨가된 음성 신호의 가청 스펙트럼과 깨끗한 음성의 가청 스펙트럼을 다음의 표현들을 사용하여 $A_y(k,i)$와 $A_x(k,i)$로 정의하자.

$$A_y(k,i) = \max\{ Y_p(k,i), T(k,i)\} \quad 0 \le k \le K-1 \tag{4}$$

$$A_x(k,i) = \max\{ X_p(k,i), T(k,i) \quad 0 \le k \le K-1 \tag{5}$$

그러므로, $A_d(k,i)$로 정의된 추가된 잡음의 가청 스펙트럼은 다음과 같이 표현될 수 있다[5].

$$A_d(k,i) = A_y(k,i) - A_x(k,i),\quad 0 \le k \le K-1 \quad (6)$$

2,3 잡음 제거를 위한 사이코어쿠스틱 기준

결국, $Y_p(k,i)$가 향상된 음성의 파워 스펙트럼 $\hat{X}_p(k,i)$을 유도하기 위해 적당하게 변형되어진다면, 그때 변형된 가청 잡음 스펙트럼 $\hat{A}_d(k,i)$는 다음을 만족시켜야 한다.

$$\hat{A}_d(k,i) \le 0, \qquad 0 \le k \le K-1 \quad (7)$$

$Y_p(k,i)$의 효과적인 스펙트럴 변형은 몇 가지 방법들에 의해 가능할 수 있다. 그러나, 선형 잡음 압축을 이용하는 기술들의 이득 곡선들은 순간적인 SNR의 함수로서 높은 순간적인 SNR에서는 어느 정도 일정하지만 낮은 일시적인 SNR에서는 심한 변화를 보인다. 그래서 파라미터 비선형 함수를 사용하여, 이득 제어에서 훨씬 더 좋은 유연성을 주었다. 이 함수는 다음과 같이 주어진다.

$$\hat{X}_p(k,i) = \frac{Y_p^{\,v(k,i)}(k,i)}{a^{\,v(k,i)}(k,i) + Y_p^{\,v(k,i)}(k,i)}\, Y_p(k,i) \quad (8)$$

식(8)에서 보듯이, 향상된 파워 스펙트럼은 양수로 가정된 두 파라미터 a(k,i)와 v(k,i)에 의해 제어되어진다. 파라미터 a(k,i)는 이 값 아래에서는 모든 주파수 성분들이 아주 잘 압축되어지는 한계값이다. 파라미터 v(k,i)는 압축하는 비율을 제어한다[4].

2.4 사이코어쿠스틱 변형을 위한 파라미터 추정

모든 스펙트럴 성분들 k에 대하여 파라미터 a(k,i)와 v(k,i)를 추정하는 것은 바람직하지 않다. 왜냐하면, 이런 방법에서 추정은 특정한 스펙트럴 값들에 매우 민감하기 때문이다. 이 때문에 하나의 특수한 주파수 영역에 대하여 a(k,i)와 v(k,i)의 고정된 값을 사용하는 것이 바람직하다. 위 과정은 CB b의 최대와 최소 범위로 k_{hb}, k_{lb}를 가진 신호의 특수한 대역에 적용된다. 이 주파수 영역에서는 a(k,i)와 v(k,i)는 $a_b(i)$와 $v_b(i)$로 정의된 상수일 것이다. 또 $v_b(i)$는 이 밴드안에서 임의의 양수값으로 두자. 향상된 파워 스펙트럼은 식(9)를 만족해야 한다.

$$\begin{aligned} &\hat{X}_p(k,i) - X_p(k,i) \le 0, \\ &\quad \text{if } Y_p(k,i) \ge T(k,i) \text{ and } X_p(k,i) \ge T(k,i) \\ &\hat{X}_p(k,i) - T(k,i) \le 0, \\ &\quad \text{if } Y_p(k,i) \ge T(k,i) \text{ and } X_p(k,i) < T(k,i) \\ &\qquad 0 \le k \le K-1 \end{aligned} \quad (9)$$

식(8)을 대입한 식(9)의 해답은 식(10)과 같다.

$$\begin{aligned} a_{Ib}(i) &= Y_p(k_I,i)\left[\frac{Y_p(k_I,i)}{X_p(k_I,i)} - 1\right]^{1/v_b(i)}, \\ &\quad \text{if } X_p(k,i) \ge T(k,i) \\ a_{IIb}(i) &= Y_p(k_{II},i)\left[\frac{Y_p(k_{II},i)}{T(k_{II},i)} - 1\right]^{1/v_b(i)}, \\ &\quad \text{if } X_p(k,i) < T(k,i) \\ &\qquad k_{lb} \le k \le k_{hb} \end{aligned} \quad (10)$$

최종적으로 $a_b(i)$는 $a_{Ib}(i)$와 $a_{IIb}(i)$ 중에서 큰 값이다[4].

$$a_b(i) = \max\{\, a_{Ib}(i),\ a_{IIb}(i) \,\} \quad (11)$$

2.5 사이코어쿠스틱 음성 향상

앞에 묘사된 파라미터 음성 향상 접근은, 특히 낮은 SNR에서는 쉽게 추정할 수 없는 깨끗한 음성 스펙트럼의 좋은 추정에 의존한다는 단점을 가지고 있다. 이 이유 때문에 이제, 완전한 음성 스펙트럼을 추정해야 되는 요구에서 단지 CB당 $X_p(k,i)$성분 중 하나의 값만 요구되는 스파스 음성 추정이 소개되어진다.

이 논문에서 제시하는 스파스(sparse) 데이터를 얻는 방법은 부분적인 최소값 성분인 $X_p(k_I, i)$(AMT 이상의 성분들 중에서) 대신에 특정한 CB에서 $X_{pb,\min}(i)$으로 정의된, 최소값 음성 파워 스펙트럼 성분을 이용하여 $a_b'(i)$를 추정할 수 있다.

$$a_b'(i) = [\, D_{pb} + X_{pb,\min}(i)\,]\left[\frac{D_{pb}}{X_{pb,\min}(i)}\right]^{1/v_b(i)} \quad (12)$$

$$X_{pb,\min}(i) = \min_k \{\, X_p(k,i),\ k_{lb} \le k \le k_{hb} \,\}$$

결국, 기존에는 깨끗한 음성 신호의 완전한 추정이 필요한데 반해 이 방법은 CB당 하나의 최소값 주파수 성분의 정확한 추정값만 있으면 되는 것이다. 그러나, 만약 최소값 주파수 성분을 정확하게 추정하지 못한다면 음성 향상은 되지 않을 것이다. 즉 최소값 주파수 성분의 정확한 추정은 상당히 중요한 문제이다. 그래서 본 논문에서는 이를 위해 변형된 스펙트럴 서브트랙션 방법을 이용하여 정확한 최소값 주파수 성분을 추정하였다[4].

3. 실험 및 결과

3.1 실험

사용된 음성 신호는 8000Hz의 샘플링 rate를 가진 디지털화된 신호이다. 그리고 잡음은 백색 가우시안 잡음(white Gaussian noise)을 사용하였다.

3.1.1 변형된 스펙트럴 서브트랙션(α,β추출과정)

본 실험의 최종 결과물로 나오는 향상된 음성의 SNR 값이 높은 향상을 가져오기 위해서는 CB당 스펙트럴 최소값의 정확한 추정이 이루어져야 한다. 이것을 위해 여기서는 먼저 식(2)에서 α는 1~6 사이의 범위에서 0.3의 간격으로 β는 0~0.5 사이의 값으로 0.02의 간격으로 변화시킴으로서 α와 β값에 따라 변하는 향상된 음성의 SNR값이 최고가 되는 곳에서의 파라미터 α와 β를 찾았다[6]. 그래서 찾은 α와 β값을 식(2)에 대입하여 얻은 G(f)를 구했다. 그리고, 식(1)과 결합해서 변형된 스펙트럴 서브트랙션의 결과 신호를 찾았다. 그런 후 그 신호를 CB당 신호 레벨을 위한 입력으로 결국 CB당 스펙트럴 최소값을 구하는데 이용했다.

3.1.2. CB당 스펙트럴 최소값의 추정

음성 스펙트럼의 최소값을 모델링은 다음과 같다.

$$\begin{aligned} \hat{X}_{b,\min} &= \frac{1}{\sqrt{2}}\sqrt{\left(\frac{1}{1+Z_{b,post}}\right)\left(\frac{Z_{b,prio}}{1+Z_{b,prio}}\right)} \\ &\cdot M\left[-\sqrt{2}\sqrt{(1+Z_{b,post})\left(\frac{Z_{b,prio}}{1+Z_{b,prio}}\right)}\right]\overline{X}_b \end{aligned} \quad (13)$$

$$Z_{b,post} \equiv \frac{\overline{X}_b^{\,2}}{\lambda_{b,\overline{X}}} - 1, \quad Z_{b,prio} \equiv \frac{\lambda_{b,\min}}{\lambda_{b,\overline{X}}}$$

이런 과정으로 추정된 CB당 스펙트럴 최소값들을 식(12)에 대입해서 파라미터 $a_b'(i)$를 구한 다음 식(8)에 대입하여 향상된 음성 신호를 구했다. 이때 파라미터 v는 1로 두었다[4].

3.2 결과

8000Hz의 샘플링 rate를 가진 "아니오"(그림 1)라는 음성신호에 크기가 각각 다른 잡음을 첨가해 SNR이 -3.83dB인 잡음이 첨가된 신호(그림 2)를 만들었다.

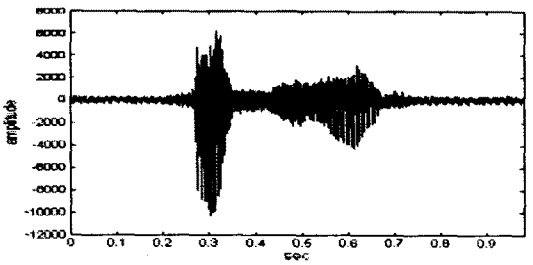

그림 1. 잡음이 첨가되지 않은 음성 신호 "아니오"

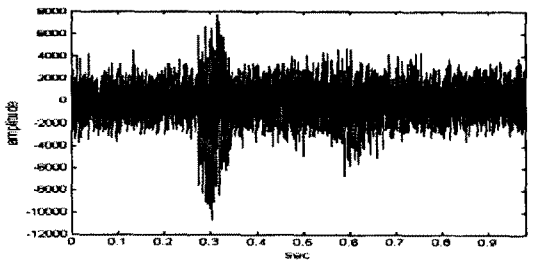

그림 2. 잡음이 첨가된 음성 신호(SNR=-3.83dB)

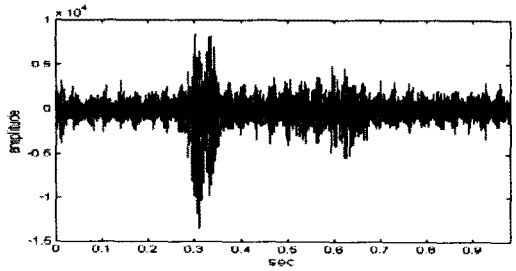

그림 3. PSS에 의한 음성 향상(SNR=-0.74dB)

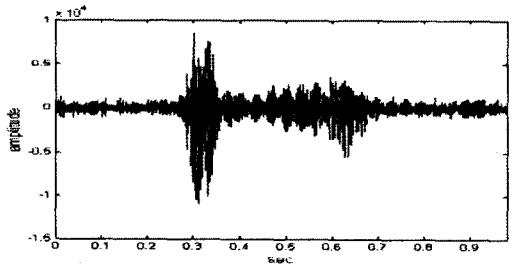

그림 4. PSS+"minima" 파라미터 방식(SNR=3.25dB)

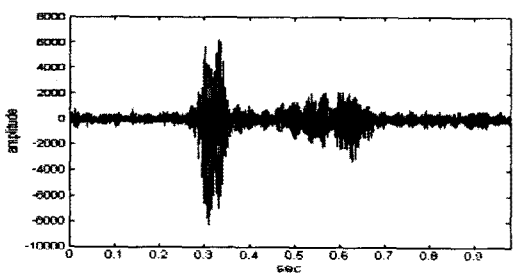

그림 5 MSS+"minima"파라미터 방식(SNR=5.08dB)

표 1. 변형된 스펙트럴 서브트랙션에서 구한 파라미터들(α, β)

parameter \ SNR[dB]	2.53	-1.08	-3.83	-4.72
α	3.20	1.20	2.80	1.30
β	0.30	0.22	0.28	0.26

표 2. 음성 향상 방법들의 SNR 비교

음성 향상의 방법 \ SNR[dB]	2.53	-1.08	-3.83	-4.72
PSS	3.68	1.42	-0.74	-1.35
PSS+"minima" parameters	6.13	4.92	3.25	3.61
MSS+"minima" parameters	7.66	6.40	5.08	4.84

4. 결 론

본 논문에서는 음성 향상을 위해 사이코어쿠스틱스 모델에 근거하여 널리 알려진 청각 매커니즘인 잡음 매스킹을 이용하였다.

음성 향상을 위해 먼저 가청 잡음의 제거가 가능한 모디파이어를 유도했다. 이러한 변형으로 지각적으로 중요한 스펙트럴 값들에게 선택적으로 영향을 주어 전체적으로 STSA에 영향을 주는 기존의 방법들보다 왜곡들이 덜 나타나게 하였고, SNR이 높은 경우 뿐만이 아니라, SNR이 낮은 경우에도 높은 향상의 결과를 보였다.

그리고, 제안된 방법은 잡음이 첨가된 음성신호로부터 모든 스펙트럴 값들 중에서 CB당 스펙트럴 최소값만으로 이용하기 때문에 정확한 추정이 필요했다. 이를 위해, 음성 향상을 위해 오래전부터 사용되어 온 파워 스펙트럴 서브트랙션의 좀 더 유연한 형태인 변형된 스펙트럴 서브트랙션을 이용해 정확한 CB당 스펙트럴 최소값을 구한 결과 음성 신호의 SNR이 향상되었다

[참 고 문 헌]

[1] H. Frazier, S. Samsam, L. D. Braida, and V. Oppenheim "Enhancement of speech by adaptive filtering," in Proc. IEEE ICASSP,pp. 251-253. Apr. 1976,

[2] W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," J, Acoust. Soc. Amer., vol. 60, pp. 911-918, Oct. 1976.

[3] S. Lim, "All pole modeling of degraded speech, IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp. 197-210, June 1978.

[4] Dionysis E, Tsoukalas,John N. Mourjopoulos, and George Kokkinakis, "Speech enhancement based on audible noise suppression," IEEE Trans. Speech, Audio Processing, vol. 5, No. 6, Nov. 1997.

[5] E. Zwicker and H. Fastl, Psychoacoustics, Facts and Models. New York: Springer-Verlag, 1990.

[6] Nathalie Virag, "Single channel speech enhancemen based on masking properties of the human auditory system," IEEE Trans. Speech, Audio Processing, vol. 7, No. 2, Mar. 1999.