2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

효율적인 잡음억제를 위한 Soft Decision 기반의 음성향상 기법

임형근, 김유진, 정재호
인하대학교 전자공학과, 디지털 신호 처리 연구실

Speech Enhancement Based on Soft Decision for Effective Noise Suppression

Hyoung-Keun Lim, Yu-Jin Kim, and Jae-Ho Chung
Digital Signal Processing Lab., Electronic Engineering, Inha University
#253, YoungHyun-Dong, Nam-Ku, InChon, Korea, 402-751
e-mail : g1991174@inhavision.inha.ac.kr

요약

비상관적인 가산잡음에 오염된 음성으로부터 향상된 음성을 얻기 위한 방법 중 Soft Decision에 근거한 음성 향상 기법이 뛰어난 성능을 가진다고 알려져 있다. Soft Decision은 주파수 영역에서 음성에 가산된 잡음을 처리하며, 잡음 환경에 대한 사전정보에 의존적이다. 본 연구에서는 Soft Decision을 근거로 음성에 가산된 잡음신호를 비선형 처리를 하여 효과적으로 음성에 포함된 잡음을 추정하도록 하였으며, 잡음환경에 대한 사전 정보 없이 효율적으로 잡음을 억제하는 방법을 제안한다. 본 연구에서 제안한 음성향상 기법은 주관적인 음질평가에서 기존의 방법들보다 나은 성능을 나타내었다.

1. 서론

잡음이 가산된 음성신호에서 그 잡음의 억제에는 음성의 왜곡이 없는 한도에서의 최대한의 잡음의 억제가 중요하다고 할 수 있다. 최근까지 비상관적인 가산잡음에 오염된 음성으로부터 향상된 음성을 얻기 위한 방법에는 Spectral Subtraction[1], Wiener Filtering[2], Soft Decision Estimation[3], 그리고 Minimum Mean Square Error Estimation[4] 등을 들 수 있으며, 그 중 Soft Decision에 근거한 방법이 뛰어난 성능을 나타낸다고 알려져 있다. 그러나 기존의 Soft Decision Estimation 방법은 음성신호와 잡음의 사전정보를 알아야 효과적인 잡음억제를 할 수 있다. 본 연구에서는 기존의 방법들 중 가장 간단한 형태의 Amplitude Estimator로 알려진 Spectral Subtraction과, 성능이 뛰어나다고 알려졌으며 확률적인 척도를 포함하고 있는 Soft Decision Estimation을 소개한다. 그리고, Soft Decision을 근거로 하여 음성에 가산된 잡음신호를 비선형 처리를 하여 효과적으로 잡음을 추정하도록 하였으며, 잡음환경에 대한 사전정보 없이 효율적으로 잡음을 억제하는 방법을 제안한다.

2. Spectral Subtraction Estimation

잡음신호 $n(t)$이 음성신호 $s(t)$에 가산된 형태를 식(1)과 같이 가정하여 표현할 수 있다.

$$y(t) = s(t) + n(t) \tag{1}$$

식(1)에서 퓨리에 변환을 통해 주파수 축으로 변환이 식(2)와 같이 되며,

$$Y(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \tag{2}$$

식(2)에서 비음성구간 동안 계산되어진 잡음의 크기인 $|N(e^{j\omega})|$의 평균을 $\mu(e^{j\omega})$라 하면, Spectral Subtraction Estimator, $\hat{S}(e^{j\omega})$는 아래와 같이 표현된다.

$$\hat{S}(e^{j\omega}) = [|Y(e^{j\omega})| - \mu(e^{j\omega})] \tag{3}$$

또는

$$\hat{S}(e^{j\omega}) = H(e^{j\omega}) Y(e^{j\omega}) \quad (4)$$

가산된 잡음은 주파수대역별로 비음성 구간에서 계산되어진 잡음 크기의 평균을 이용하여 음성에 포함된 잡음을 차감하는 방법으로 제거된다.

$$H(e^{j\omega}) = 1 - \frac{\mu(e^{j\omega})}{|Y(e^{j\omega})|} \quad (5)$$

$$\mu(e^{j\omega}) = E(|N(e^{j\omega})|) \quad (6)$$

일반적으로 Spectral Subtraction Estimation은 구현이 용이하고 간단하지만 음성신호와 잡음의 크기에 따라 음성이 왜곡될 가능성이 크며 원하지 않은 잡음(musical noise)이 발생하고, 음성의 왜곡정도가 심한 단점이 있다.

3. Soft Decision Estimation

식(1,2)에서 음성신호의 부재와 존재에 대한 가정을

음성부재: $H_0: |y_n| = |n_n|$

음성존재: $H_1: |y_n| = |A e^{j\omega} + n_n|$ (9)

와 같이 하면, 음성존재와 부재에 대한 가정을 바탕으로 Soft Decision Estimator는

$$\hat{A} = E(A|V, H_1)\, P(H_1|V) + E(A|V, H_0) P(H_0|V) \quad (10)$$

로 표현된다. 식(10)에서 $V = |y_n|$이며, $P(H_k|V)$는 음성이 존재 또는 부재할 확률이다. 식(10)에서 Soft Decision Estimator를 다시 표현하면,

$$\hat{A} = E(A|V, H_1) P(H_1|V) \quad (11)$$

이 된다. 식(11)에서 $E(A|V, H_1)$은 음성이 존재할 때 A의 MVU(Minimum Variance Unbiased) Estimator를 의미하며, ML(Maximum Likelihood) Estimator로 표현할 수 있다. 주파수 채널별 음성존재 사후 확률, $P(H_1|V)$, 은 Bayes rule을 적용하여,

$$P(H_1|V) = \frac{P(V|H_1)P(H_1)}{P(V|H_1)P(H_1) + P(V|H_0)P(H_0)} \quad (12)$$

으로 나타낼 수 있으며, $P(V|H_k)$는 각각의 음성부재 또는 존재에 대한 사전 확률밀도함수로

$$P(V|H_0) = \frac{2V}{\lambda_n} \exp\left(-\frac{2V^2}{\lambda_n}\right) \quad (13)$$

$$P(V|H_1) = \frac{2V}{\lambda_n} \exp\left(-\frac{2V^2 + A^2}{\lambda_n}\right) I_0\left(\frac{2AV}{\lambda_n}\right) \quad (14)$$

으로 정의 할 수 있다. 식(13,14)에서 λ_n은 평균 잡음 전력이며, $I_0[\cdot]$는 modified Bessel function 이다. *a priori SNR*(the suppression factor), ξ, 을

$$\xi = \frac{A^2}{\lambda_n} \quad (15)$$

로 정의하면, 음성존재의 사후 확률, $P(H_1|V)$은

$$P(H_1|V) = \frac{\exp(-\xi)\, I_0\left[2\sqrt{\xi\left(\frac{V^2}{\lambda_n}\right)}\right]}{1 + \exp(-\xi)\, I_0\left[2\sqrt{\xi\left(\frac{V^2}{\lambda_n}\right)}\right]} \quad (16)$$

으로 변형되며, 최종적으로

$$\hat{s} = \hat{A}\frac{y}{|y|} = \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{V^2 - \lambda_n}{V^2}}\right] \cdot P(H_1|V) \cdot y \quad (17)$$

으로 가산된 잡음으로부터 향상된 음성신호를 얻을 수 있다. 일반적으로 Soft Decision Estimation은 기존의 Amplitude Estimator에 음성의 존재 확률을 고려하여 보다 정확히 잡음을 억제하며 성능도 뛰어난 것으로 알려져 있다. 하지만 어느 정도 만족한 성능을 얻기 위해서는 음성신호와 잡음의 사전 정보(ξ : the suppression factor)를 알고 있어야 하며, 그 사전정보를 모른다면 음성에 가산된 잡음을 효율적으로 억제하기 어렵다. 또한 신호의 크기를 추정해 가는 Amplitude Estimator의 기반에 확률적인 척도를 도입하기 때문에 음성과 잡음의 크기의 차이에 따라 음성의 크기를 추정하기 보다 잡음의 크기를 추정해 갈 수도 있는 단점이 있다.

4. 제안된 음성향상 기법

제안된 음성향상 기법에서는 기존의 Soft Decision Estimation에서 음성신호의 부재와 존재에 대한 가정을 비선형 처리 한 상태로 식(18)과 같이 시작한다.

음성부재가정: $H_0: \log|y_n| = \log|n_n|$

음성존재가정: $H_1: \log|y_n| = \log|A e^{j\omega} + n_n|$ (18)

이와 같이 하면, 비선형 처리를 통한 가정에 의해서 잡음만이 존재하는 음성부재 가정의 경우에는 잡음의 크

기가 불규칙적으로 심하게 변하는 정도에 둔감하게 작용을 하며, 음성과 잡음이 동시에 존재하게 되는 음성존재가정의 경우에는 음성과 잡음의 경계를 보다 명확하게 구분할 수 있다. 위 가정을 바탕으로 $V=\log|y_n|$이 되며, Maximum Likelihood Estimator로 신호의 크기를 식(19)와 같이 추정할 수 있다.

$$\hat{A}=\frac{1}{2}\left[\log|y_n|+\sqrt{\log|y_n|^2-\log\lambda_n}\right] \quad (19)$$

이와 같이 추정된 신호는 Soft Decision Estimation의 사전정보인 *a priori SNR*(ξ : the suppression factor)에 식(20)과 같이 적용하여 기존의 Soft Decision Estimation에서 음성신호와 잡음의 사전정보가 없어도 가능하도록 추정된 $\hat{\xi}$ 을

$$\hat{\xi}=\frac{\hat{A}^2}{\lambda_n} \quad (20)$$

와 같이 구한다. 식(19)에서 비선형 처리를 통하여 구해진 각각의 값들은 식(20)에서 추정된 사전정보($\hat{\xi}$)와 같이 음성존재 사후 확률인 $P(H_1|V)$에 식(21)과 같이

$$P(H_1|V)=\frac{\exp(-\hat{\xi})\,I_0\left[2\sqrt{\hat{\xi}\left(\frac{2\log|y_n|}{\log\lambda_n}\right)}\right]}{1+\exp(-\hat{\xi})\,I_0\left[2\sqrt{\hat{\xi}\left(\frac{2\log|y_n|}{\log\lambda_n}\right)}\right]} \quad (21)$$

으로 되며, 최종적으로 가산된 잡음으로부터 향상된 음성신호는 변형된 Maximum Likelihood Estimator와 음성존재 사후확률, $P(H_1|V)$,로 주파수 채널별 이득(Gain Factor)를 식(22)와 같이 얻을 수 있다.

$$\hat{G}=\left[\frac{1}{2}+\frac{1}{2}\sqrt{\frac{2\log|y_n|-\log\lambda_n}{2\log|y_n|}}\right]\cdot P(H_1|V) \quad (22)$$

구해진 주파수채널별 이득을 이용하여 향상된 음성신호는 식(23)과 같이 구할 수 있다.

$$\hat{s}=\hat{G}\cdot\log|y_n| \quad (23)$$

위 수식들에서 알 수 있듯이 제안된 음성향상 기법은 $\hat{\xi}$ 을 통하여 음성신호와 잡음의 사전 정보 없이 향상된 음성을 얻을 수 있으며, log를 이용한 신호의 비선형 처리로 음성과 잡음의 크기의 차이에 따른 추정을 더욱 명확히 하여 기존의 음성향상 기법보다 나은 결과를 도출해 낼 수 있다.

5. 실험결과 및 결론

실험에는 <그림1>과 같이 8000Hz, 16-bit, PCM 데이터를 사용하였다.

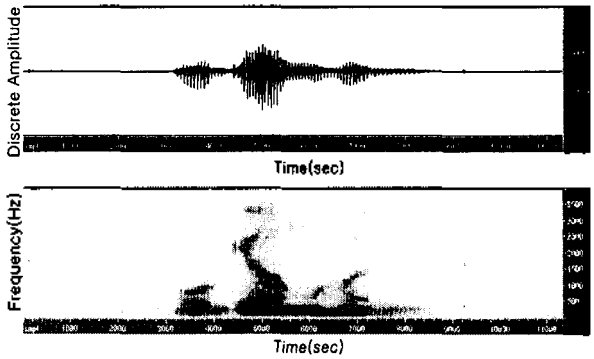

그림 1 깨끗한 음성 (발음 : "오경훈")

주어진 음성신호를 NOISEX-92 데이터 베이스 중에서 백색 배경잡음을 SNR-10dB로 가산하여 잡음이 포함된 음성데이터를 <그림2>와 같은 얻었다.

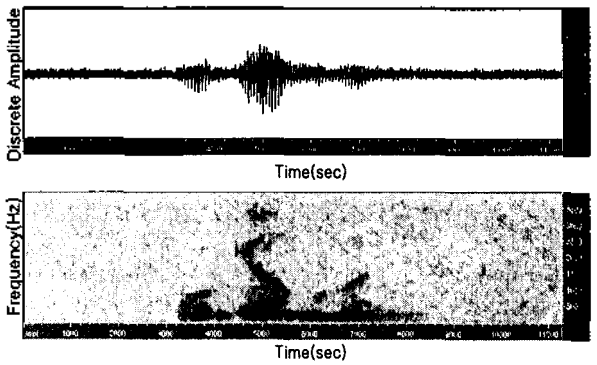

그림 2 잡음이 가산된 음성 (SNR 10dB)

기존의 Soft Decision Estimation을 수행하기 위해 잡음이 가산된 음성 신호를 32ms 크기의 윈도우를 사용하여 Half-Overlapping 하였고 기존의 Soft Decision Estimation의 <Implementation>에서 제시하는 Filter Banking과 Smoothing등을 하여 <그림3>과 같은 결과를 얻었다.

제안된 음성향상기법의 수행을 위하여 동일한 환경에서 잡음이 가산된 음성 신호를 32ms 크기의 윈도우를 이용하여 Half-Overlapping 하였고 선택된 윈도우마다 FFT를 수행하여 주파수 대역으로 변환 후 그 신호

의 크기에 비선형 처리를 하여 결과를 얻었다. 본 연구에서 제안된 음성향상 기법에서는 Filter Banking 이나 Smoothing 등을 고려하지 않은 기본적인 Soft Decision 알고리즘만을 사용하여 <그림4>와 같은 결과를 얻었다.

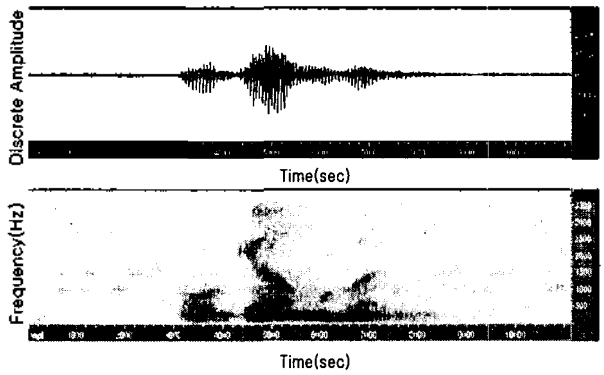

그림 3 기존의 Soft Decision 수행 결과

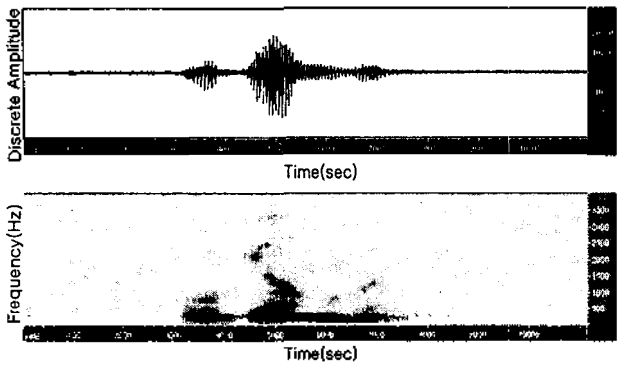

그림 4 제안된 음성향상 기법 수행 결과

기존의 Soft Decision Estimation을 이용한 결과와 제안된 음성향상 기법을 이용한 결과를 개인적으로 비교하였다. 비교되는 기준에는 첫째, 음성의 왜곡이 최소한 경우에서의 잡음이 가능한 최대로 억제되었는지, 둘째, 일반인이 듣기에 거부감이 없이 매끄러운 지의 여부 그리고 셋째, 잡음이 포함되지 않은 음성과 그 특성이 얼마나 가까운지를 판단하였다. 지금까지의 실험 과정을 통해 다음과 같은 사실들을 알게 되었다. 기존의 Soft Decision 경우 적절한 *a priori SNR*(the suppression factor), ξ, 값을 설정하기 쉽지 않았으며, 잡음이 부자연스럽게 억제되었고, 스펙트럼 상에서의 보기와 달리 음성이 왜곡되어 듣기에 부자연스러웠다. 반면에 제안된 음성향상 기법의 경우 잡음이 기존의 방법보다는 적게 억제가 되었지만 자연스러웠으며 기존의 방법보다 음성의 왜곡이 거의 없으며 깨끗한 음성의 특성에 가까웠다.

6. 앞으로의 연구방향

앞으로 이득에 대한 Smoothing과 같은 처리를 통해 제안된 음성향상 기법을 보완할 것이며 보다 객관적인 평가를 위해 여러 음소가 골고루 들어가 있는 충분한 음성 데이터를 확보하고 Babble / Pink / White Noise 등의 다양한 가산 잡음 환경에서 SNR을 달리하여 MOS 평가를 시도할 것이다. 또한 제안된 음성향상 기법을 보다 객관적으로 평가하기 위해서 상용화된 음성향상 알고리즘과의 샘플 비교 실험과 MOS 평가를 수행하고자 한다. 마지막으로 제안된 음성향상 기법이 음성인식 미치는 영향에 대해서 연구를 병행할 예정이다.

7. 참고문헌

[1] Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. on ASSP*, vol. 29, April 1979.

[2] Lim, Openheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. IEEE*, vol. 67, No. 12, Dec. 1979

[3] McAulay, Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. on ASSP*, vol. 28, NO. 2, April 1980.

[4] Ephraim, Malah, "Speech Enhancement Using MMSE Short-Time Spectral Amplitude Estimator," *IEEE Trans. on ASSP*, vol. 32, NO. 6, Dec. 1984.

[5] Yang, "Frequency Domain Noise Suppression Approaches in Mobile Systems," *ICASSP*, 1993.

[6] Scalart and Vieira Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," *ICASSP*, 1996.