1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

이중 여진 음성모델을 이용한 음질개선

유창동
한국통신 멀티미디어연구소 음성언어연구실

A Voice/Unvoice Decomposition in Noisy Background

Chang Dong Yoo
Spoken Language Research Team, Multimedia Technology Research Laboratory, Korea Telecom

Email : cdyoo@smm.kotel.co.kr

요약

음질개선에 이중 여진(Double Excitation) 음성모델을 적용하는 방법이 있다. 유성음과 무성음 성분들로 분리하는 이 방법은 각 성분들의 고유한 성질을 이용하여 음질을 저하시키는 wideband 잡음을 제거할 수 있다. 이중 여진 음성모델을 이용한 음질개선 시스템과 기존의 스펙트랄 제거(spectal subtraction) 알고리즘을 비공식적으로 비교한 결과 이중 여진 모델을 이용한 방법이 더 나은 성능을 보였다.

I. 서론

일반적으로 통신 시스템에는 음질저하를 일으키는 많은 음향 잡음이 존재하며, 그 잡음의 형태도 사무실에서 전화통화중에 생기는 작은 잡음에서부터 비행기나 헬리콥터의 큰 소음까지 다양하다. 이러한 잡음들은 시스템의 성능을 저하시키며 청취자를 피로하게 만들기 때문에 음성신호로부터 잡음신호를 줄이는 자동적인 음질개선 시스템의 개발이 필요하다.

다양한 형태의 음질개선 시스템들이 제안되어 왔으며[1, 2, 3, 4], 이러한 시스템들의 성능은 제거하고자 하는 잡음의 종류와 시스템이 필요로하는 잡음의 정보에 따라 좌우되었다. 본 연구에서는 잡음신호와 음성신호가 공존하는 단일 신호에서 wideband 잡음을 제거하는데 주안점을 두었다.

음성신호의 복잡성과 기존에 제안되어 왔던 음성 모델들의 한계 때문에 음질개선 시스템에는 음성 모델을 기초로한 방법이 거의 사용되지 않아왔다. 특히 음성 모델을 이용한 기존의 음질개선 시스템들은 신호대 잡음비(signal-to-noise ratio)를 떨어뜨리는 왜곡현상으로 음성신호를 변질시켜 시스템의 성능을 저하시켰다. 결과적으로 기존의 대부분의 음질개선 시스템들은 기본적인 음성모델에 기초를 두지 않고 음성 파형을 직접 이용한 시도만 해왔다.

가장 널리 쓰이는 음질개선 방법은 스펙트랄 제거 방법으로, 이 방법의 기본적인 원리는 신호대 잡음비가 낮은 부분의 주파수를 약화시키는 것이다. 반면에 신호대 잡음비가 높은 부분에서는 주파수를 상대적으로 약간만 줄인다. 스펙트랄 제거 방법은 신호대 잡음비가 클 경우 효과적으로 사용될 수 있지만 이 방법을 이용한 잡음감소 효과는 기본적인 음성의 인지도를 떨어뜨린다. 적절한 양의 잡음감소는 음성의 인지도를 크게 떨어뜨리지 않겠지만 큰 폭으로 잡음을 감소시킬 경우에는 음성의 인지도를 심각하게 악화시킬 수 있다. 이러한 스펙트랄 제거방법은 특히 고주파 형태의 무성음을 악화시킬 수 있다. 이런 성질이 음성의 인지도를 약화시키는 주 원인이 된다. 스펙트랄 제거 방법에 의해서 생기는 또 다른 왜곡현상은 잡음제거를 거친 음성신호가 음색 잡음(tonal noise)을 포함하게 된다는 것이다.

본 논문에서는 위에서 설명한 문제점들을 해결할 수 있는, 이중 여진 음성 모델에 기초한, 새로운 음질개선 시스템을 소개하고자 한다. 이중 여진 음성 모델은 음성을 유성음과 무성음으로 분리하는데 이용된다. 여러 경우에 배경으로 깔리는 음향 잡음은 무성음과 유사한 특성을 가지고 있기 때문에 분리된 무성음 신호부분은 음성의 무성음과 배경잡음의 합으로 이루어질 것이다. 유성음 부분은 주기성을 띤 음성신호로 구성된다. 결과적으로 무성음 신호부분에서 현저한 잡음을 감소시킴으로써 잡음이 제거될 것이다. 새롭게 제안된 이 방법은 잡음제거를 거친 후 생기는 왜곡현상을 감소시키기 위하여 유성음 성분과 무성음 성분으로 분리된 각 부분의 특성을 효과적으로 이용함으로써 진행된다.

II. 이중 여진 음성 모델

이중 여진 음성 모델은 기존의 음성 모델의 한계점을 보완할 수 있다. 기존의 모델은 음성을 여진 신호의 나열을 입력으로 한 선형필터의 응답으로 보고, 그 여진 신호의 특성에 따라 유성음 또는 무성음으로 모델을 정한다. 유성음은 주기적인 임펄스(Impulse) 나열로 무성음은 백색 잡음 나열로 여진 신호를 모델링 한다. 이러한 음성 모델은 유성음과 무성음을 구분하는데 많은 어려움이 따르고 특히 원음에 잡음이 있을 때 음성을 분석하는데 더 큰 어려움이 따른다. 모델을 기초로 음성을 분석하고 추정하는 알고리즘들은 음질의 손상을 복원시키는데 안정적이지 못하다. 이중 여진 음성 모델에서는 음성신호 $s_w(n)$을 독립적인 두 성분, 유성과 무성음 성분으로 분리한다. 각 성분은 각각 $v_w(n)$과 $u_w(n)$으로 표시한다. 청자는 윈도우 함수 $w(n)$에 의해서 구해지는 한 구간을 나타낸다. 음성 신호 $s_w(n)$은 Fourier 도메인에서 다음과 같은 식으로 표현할 수 있다.

$$S_w(\omega) = V_w(\omega) + Uw(\omega) \qquad (1)$$

$S_w(\omega)$, $V_w(\omega)$, $U_w(\omega)$은 각각 $s_w(n)$, $v_w(n)$, $u_w(n)$의 Fourier 변환을 나타낸다.

유성음 성분은 정의에 따라 윈도우 $w(n)$ 구간에 대해서 주기성을 가지는 것으로 가정한다. 그러므로 각 음성구간의 유성음을 나타내는 부분은 윈도우를 주기 P_0에 따라 하모닉 모듈(harmonic modulation)했다고 본다. 유성음 성분을 나타내는 $v_w(n)$과 $V_w(\omega)$은 아래 식과 같다.

$$v_w(n) = \sum_{m=-M}^{M} A_m w(n) e^{jm\omega_0 n} \tag{2}$$

$$V_w(n) = \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \tag{3}$$

$W(\omega)$은 윈도우 함수 $w(n)$의 Fourier 변환이며 기본적으로 대역폭이 좁은 로우패스(lowpass) 필터이다. 그러므로 $V_w(\omega)$은 기본 주파수 ω_0의 다양한 harmonic의 합으로 나타내어 진다. A_m은 m번째 주기함수의 진폭 크기를 나타내는 계수이다. ω_0는 기본적인 주파수로 다음식과 같이 피치 주기 P_0로 표현된다.

$$\omega_0 = \frac{2\pi}{P_0} \tag{4}$$

harmonic의 계수 M은 ω_0의 함수로 다음과 같다. $\lfloor \bullet \rfloor$는 매개변수 값 이하의 가장 큰 정수를 나타낸다.

$$M = \left\lfloor \frac{\pi}{\omega_0} \right\rfloor \tag{5}$$

실제로 이중 여진 모델의 계수값들은 알려져 있지 않기 때문에 음성 스펙트럼을 이용하여 추정하여야만 한다. 이렇게 추정된 기본 주파수, 주기함수의 진폭, 유성음 스펙트럼은 각각 $\hat{\omega}_0$, $\hat{A}_m$, $\hat{V}_w$로 표기한다. 기본 주파수와 주기함수의 진폭의 추정값은 Griffin[6]이 개발한 알고리즘을 이용하여 구한다. 이 알고리즘은 음성 스펙트럼 $S_w(\omega)$과 유성음 스펙트럼 $V_w(\omega)$ 사이의 mean-squared error를 최소화 하며, 유성음 성분이 원음의 harmonic 성분을 띤 부분을 포함한다는 것을 보장한다. 무성음 스펙트럼 $U_w(\omega)$은 음성 성분에서 유성음 성분을 뺀 스펙트럼 $D_w(\omega)$을 이용해서 추정한다.

$$D_w(\omega) = S_w(\omega) - \hat{V}_w(\omega) \tag{6}$$

$D_w(\omega)$를 이용해서 $U_w(\omega)$를 추정해내는데 다양한 방법들이 있다[5]. 이러한 방법들은 $D_w(\omega)$ 스펙트럼을 smoothing 하는데 있어서 여러가지 형태의 방법들을 허용하는데 그 이유는 무성음 스펙트럼의 원 구조를 보존할 필요가 없기 때문이다. 무성음 스펙트럼은 $D_w(\omega)$를 smoothing하여 공존하는 잡음을 감소시켜 구하고 무성음의 위상은 $D_w(\omega)$에서 구하거나 잡음이 섞인 음성신호에서 구한다. 그림 1의 (a)는 "버스"라고 말한 음성신호이다. 이 음성신호를 유성음 성분과 무성음 성분으로 분리한 것이 각각 그림 1의 (b)와 (c) 이다.

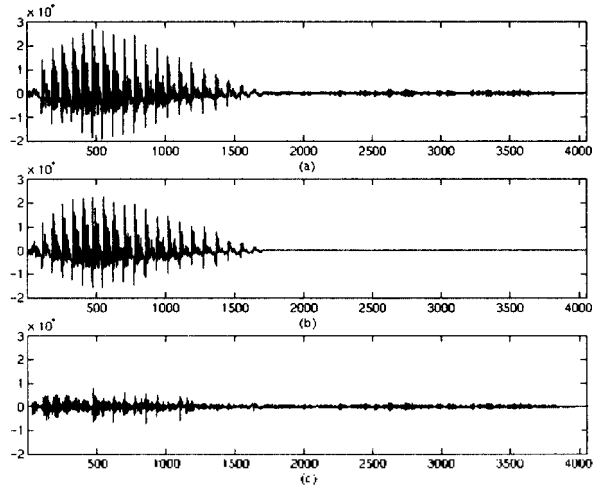

그림 1. 유/무성음 성분의 분리

III. 새로운 음질개선 방법

그림 2는 이중 여진 음질개선 시스템의 구성도를 나타낸다. 이 시스템은 유성음과 무성음 부분을 분리하여 음질을 개선시킨다. 유성음 성분은 harmonic에 포함되어 있는 잡음성분을 개선시키기 위하여 소폭의 변형만 필요하며, 이중 여진 음질개선 시스템의 대부분의 기능은 무성음 성분을 개선시키는 것이다.

우성음 성분은 주기성을 띠고 있지 않기 때문에 원음의 기본적인 음질을 왜곡시키지 않고 smoothing할 수 있다. 무성음 성분을 smoothing 하게되면 보다 좋은 무성음 성분의 추정치를 얻을 수 있는 잇점이 있다. 이 추정치의 질은 다음 단계의 스펙트랄 제거에 매우 중요한 영향을 미친다.

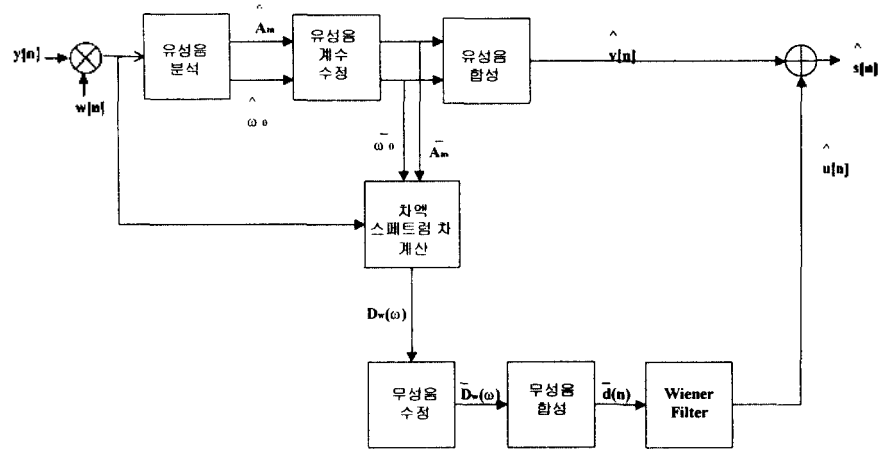

그림 2. 이중 여진 음질개선 시스템 구성도

3.1 유성음 성분의 음질개선

음성에 포함되어 있는 잡음은 결과적으로 추정된 계수 치들이 잡음요소를 내포하게 한다. 유성음 성분은 harmonics series 로 정의되기 때문에 유성음 성분에 내포되어 있는 잡음요소도 주기성을 띠게된다. 유성음 성분의 특징을 나타내는 기본요소로 두 개의 계수가 있다. 하나는 기본 주파수이고 다른 하나는 harmonic 의 진폭값이다. 기본 주파수 추정치의 오차는 무시할만하다고 가정한다[5]. 그렇기 때문에 유성음 성분의 음질개선은 주기함수들의 진폭을 수정함으로써 유성음 성분의 잡음요소를 감소시키게 된다. 잡음요소를 포함하고 추정된 진폭값 $\hat{A}_m$은 잡음요소를 제거할 수 있도록 조정되어야만 한다. m 번째 hharmonic 의 진폭 추정치 $\hat{A}_m$은 상응하는 주파수에 존재하는 실질적인 잡음요소 보다 작을경우 삭제된다. $\bar{A}_m$ 은 $\hat{A}_m$의 개선된 값을 나타낸다.

$$\bar{A}_m = \begin{cases} 0 & if\ |A_m| < 3\left[P_{zz}(m\omega_0)/N_{eff}\right] \\ A_m & otherwise \end{cases} \quad (7)$$

$P_{zz}(\omega)$는 잡음의 power density 를 나타내고 N_{eff}는 다음 식과 같이 정의되는 윈도우 효과를 나타낸다.

$$N_{eff} = \frac{[\sum_{n=-\infty}^{\infty} w^2(n)]^2}{\sum_{n=-\infty}^{\infty} w^4(n)} \quad (8)$$

위 식 (7)을 이용하여 음성이 누설된 부분은 실질적인 유성음 성분을 손상시킬 수 있지만, 일반적으로 이 부분에서는 신호대 잡음비가 낮기 때문에 손실이 크게 감지되지 않는다. 그러나 유성음 성분에서 제거된 에너지를 복원하기 위해서는 반드시 차액 스펙트럼을 수정하여야 한다.

3.2 무성음 성분의 음질개선

이 부분의 목적은 무성음 성분에서 가능한 최대로 잡음 z(n)을 제거하는 것이다. 잡음제거는 두 단계 음질개선 시스템(two-pass enhancement system)에 의해서 이루어 진다[5]. 첫 단계는 주기성 구간으로 유성음 성분의 에너지가 무성음 성분보다 현저하게 큰 부분에서는 무성음 에너지가 유성음 에너지에 의해 가려져 인지된 음성의 변형이 없이도 무성음 에너지가 제거될 수 있다. 그렇지 않을 때는 그대로 둔다. 차액 스펙트럼 $D_w(\omega)$에서 잡음을 제거한 것을 $\bar{D}_w(\omega)$로 표시하며 다음 식과 같다.

$$\bar{D}_w(\omega) = \begin{cases} 0 & if\ |E_{v_m}| > 3E_{uv_m} \\ D_w(\omega) & otherwise \end{cases} \quad (9)$$

E_{v_m}과 E_{uv_m}은 m번째 구간에서 유성음 성분과 무성음 성분의 각 에너지를 나타낸다. 두 번째 단계는 대부분의 잡음제거가 수행되는 부분으로 $\bar{D}_w(\omega)$로부터 합성된 $\bar{d}(n)$을 수정된 Wiener 필터 $\hat{H}_{wss}(\omega)$에 적용시킨다. 필터는 신호대 잡음비가 낮은 구간의 무성음에서 배경잡음을 제거한다. 이 필터의 식은 다음과 같다.

$$\hat{H}_{wss}(\omega)=\begin{cases}\beta & if\ \dfrac{\alpha E[|Z_{w_{ss}}(\omega)|^2]}{E[|D_{w_{ss}}(\omega)|^2]}>1 \\ 1.0-\dfrac{\alpha E[|Z_{w_{ss}}(\omega)|^2]}{E[|D_{w_{ss}}(\omega)|^2]} & otherwise\end{cases} \quad (10)$$

$E[|Z_{w_{SS}}(\omega)|^2]$는 잡음의 파워 스펙트럼을 나타내며 $E[|D_{w_{SS}}(\omega)|^2]$는 smoothing된 무성음 성분의 파워 스펙트럼을 나타낸다. 첨자는 윈도우 함수 $w_{ss}(n)$에 의해서 구해지는 작은 구간을 나타낸다. 상수 α와 β의 값은 각각 1.6과 0.1 이다.

4. 실험 결과

잡음이 포함된 여러 개의 음성을 가지고 위에서 설명한 이중 여진 음질개선 시스팀을 시험해 보았다. 이 음성들은 깨끗한 원음성에 Gaussian 잡음을 추가시켜 생성하였다. 이 음성들의 신호대 잡음비는 10 ~ 30 dB 이다. 이 음성들을 이용하여 기존의 스펙트랄 제거와 이중 여진 음질개선 시스팀에 대해서 각각 시험해 보았고 비공식적인 음질의 평가로 성능을 비교했다. 비교결과 이중 여진 잡업제거 시스팀의 성능이 더 우수하였다. 이중 여진 음질개선 시스팀에 의해서 처리된 음성에 인공적인 왜곡이나 음색변화가 거의 없었고, 스펙트랄 제거 방법 보다 잡음도 훨씬 더 감소되었다.
청각 장애자들을 위해 음질개선 시스팀을 이용하기 위한 연구가 M.I.T. 전자연구소와 William Rabinowitz 박사에 의해서 수행되었다. 청각 장애자들에게 음질이 저하된 음성과 음질개선 시스팀으로 처리한 음성을 들려주었다. 남성 음성의 경우 음질이 저하된 음성에 비해 이중 여진 음질개선 시스팀으로 처리한 음성의 인지도가 15% 향상되었고, 여성의 목소리의 경우에도 유사하게 인지도가 23% 향상되었다.

5. 결론

본 논문에 이중 여진 음질제거 시스팀과 그 성능평가에 대해서 기술하였다. 비공적인 청취자들을 대상으로한 실험에서 새롭게 제안한 시스팀이 기존의 스펙트랄 제거 시스팀에 비해서 성능이 우수하였다. 기존의 방법과 비교해서 감소된 잡음의 양은 유사하지만 기존의 스펙트랄 제거 시스팀에서 존재했던 인공적인 음색의 왜곡이 이중 여진 음질제거 시스팀에서는 없었다. 이러한 결과들은 이중 여진 음질제거 시스팀을 이용할 경우 청각 장애자들이 잡음이 포함된 저하된 음성을 듣는데 인지도가 향상될 것임을 보여준다.

6. 참조문헌

[1] B. Widrow and et. al., "Adaptive noise cancelling:Principles and applications," Proceedings of the IEEE, vol. 63, pp. 1692-1716, December 1975.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. on Acoustic, Speech and Signal Processing, vol. ASSP-28, pp. 137-145, April 1980.

[3] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Trans. on Acoustic, Speech and Signal Processing, vol. ASSP-26, pp. 471-472, October 1978.

[4] R. S. M. Berouti and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc., pp. 208-211, April 1979.

[5] J. Hardwick, The Dual Excitation Speech Model. PhD thesis, MIT. E.E.C.S. Department, June 1992.

[6] D. W. Griffin and J. Lim, " Anew pitch estimation algorithm," Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc., vol. 67, pp. 592-601, March 1984