VoIP 환경에서의 잡음제거를 위한 최적화된 위너 필터

정상배(ICU), 이성독(ICU), 한민수(ICU)

<차 례>

<Abstract>

Optimized Wiener Filter for Noise Reduction in VoIP Environments

Sangbae Jeong, Sungdoke Lee, Minsoo Hahn

Noise reduction technologies are indispensable to achieve acceptable speech quality in VoIP systems. This paper proposes a Wiener filter optimized to the estimated SNR of noisy speech for the noise reduction in VoIP environments. The proposed noise canceller is applied as a pre-processor before speech encoding. The performance of the proposed method is evaluated by the PESQ in various noisy conditions. In this paper, the proposed algorithm is applied to G.711, G.723.1, and G.729A which are all VoIP speech codecs. The PESQ results show that the performance of our proposed noise reduction scheme outperforms those of the noise suppression in the IS-127 EVRC and the ETSI standard for the advanced distributed speech recognition front-end.

* Keywords: Noise reduction, Speech enhancement, Voice over IP, Speech codec.

1. 서 론

음성의 품질은 주변 잡음이 존재할 경우에 심각하게 떨어지게 된다. 따라서, 음질 개선 및 잡음 제거 기술은 여러 가지 음성 인터페이스 분야에서 매우 필수적이라고 말할 수 있다. 음성에 섞인 잡음을 제거하기 위해서 주파수 스펙트럼 차감법, 위너 필터, 칼만 필터 등의 적용 신호처리 기법이 제안되어 왔다 [1]-[3][5][8]-[11][14]. 이러한 기술들은 음성 통신, 음성 인식, 음원 위치 추적 등의 다양한 분야에 적용이 되고 있다.

음성 통신 영역에서 잡음 제거 기법이 사용될 수 있는 분야로는 원격 화상 회의, Voice-over-IP (VoIP) 서비스, 유무선 망을 통한 개인 휴대 단말 통신 등을 들 수 있다. 그 중에서 VoIP 서비스는 급속하게 발전하고 있는데 통화 비용 절감 효과가 그것의 가장 큰 이유이다. 배경 잡음은 VoIP의 Quality-of-Service (QoS)를 떨어뜨리는 중요한 요인이다. VoIP 시스템에서 음성 신호는 일반적으로 한 개의 마이크로 수신되기 때문에 위너 (Wiener) 필터 및 칼만 (Kalman) 필터 등의 단일 마이크 기반의 잡음제거 기술만이 사용될 수 있다. 음성통신 시스템에서는 음질 뿐만 아니라 알고리즘 및 시스템에 의한 시간 지연 역시 QoS 측면에서는 중요한 요소이기 때문에 시간 지연을 크게 요구하는 잡음 제거기는 바람직하지 못하다. VoIP 환경에서 송수화자간의 음성 신호 지연의 요소는 전송단의 신호 버퍼링 지연, 인코딩 지연, 네트워크 지연, 네트워크 버퍼링 지연, 수신단의 디코딩 지연으로 크게 나눌 수 있다[12]. 송수화자간의 신호 지연이 어떤 임계치를 넘어서면 통화의 부자연성에 따른 불편함이 유발되는데, ITU-T G.114에서는 소위 'mouth-to-year delay'가 150 ms를 넘지 않을 것을 권장하고 있다. 따라서, 음성 부호화기의 전처리기로 사용되는 잡음 제거 알고리즘도 시간 지연과 그 성능간의 trade-off를 잘 감안하여 설계되어야 한다.

하나의 마이크를 사용하는 잡음 제거 알고리즘 중에서 위너 필터 기법은 개념적으로 단순하고 안정적인 성능을 보이는 것으로 알려져 있기 때문에 여러 가지 음성 인터페이스 분야에 널리 응용되고 있다. 따라서, 본 논문에서는 위너 필터 기법을 음성에 섞이는 잡음을 제거하기 위한 기본 알고리즘을 채택하였고 VoIP 시스템을 위한 최적화 기법을 제안한다. 본 논문에서 위너 필터의 최적화는 필터의 설계에 필요한 입력 신호대 잡음비 (SNR: Signal-to-Noise Ratio)의 변형에 의해서 이루어진다. 알고리즘의 성능은 ITU-T P.862에서 제안된 perceptual evaluation of speech quality (PESQ)에 의해서 객관적으로 측정된다[7]. 테스트를 위한 가산성 잡음은 백색 잡음, 사무실 환경 잡음, 배블(babble)성 잡음, 차량 잡음 등이 고려된다. 제안된 알고리즘의 성능은 IS-127 EVRC (enhanced variable rate codec)의 잡음 제거기 및 분산 환경 음성인식을 위한 ETSI (European Telecommunications Standard Institute) 표준의 잡음 제거기 등과 비교된다[2][4]. 각 잡음 제거기 모듈은 현재

VoIP 용 음성 부호화기로 사용되고 있는 G.711, G.723.1, G.729A 등의 전처리기로 적용되어 성능이 측정된다.

논문의 구성은 2장 본론에서 위너 필터의 개념과 제안된 방식의 위너 필터 기법, 3장 실험 및 결과부에서 실험 조건 및 조건 별 성능, 4장 결론부에서 연구의 결과 정리 및 향후 연구 계획 등이 제시된다.

2. 본 론

2.1. 표준 위너 필터의 유도

표준 위너 필터의 유도에 앞서서 가산성 잡음에 의한 음성 신호의 왜곡의 과정을 식 (1)과 같이 표현할 수 있다.

$$x(n) = d(n) + v(n) \tag{1}$$

여기서, $d(n)$은 입력 음성 신호, $v(n)$은 가산성 잡음이다. 만약 음성 신호 $d(n)$과 잡음 신호 $v(n)$이 wide-sense stationary (WSS) 하고 통계적으로 독립이라고 가정할 수 있다면, 위너 필터는 식 (2)를 최소화함으로써 구현할 수 있다 [6].

$$\zeta = E[e^2(n)] = E[(d(n) - \hat{d}(n))^2] \tag{2}$$

$\hat{d}(n)$은 음성 신호의 최적 추정치이며 식 (3)의 컨벌루션으로 표현된다.

$$\hat{d}(n) = \sum_{l=-\infty}^{\infty} w(l)x(n-l) \tag{3}$$

여기서, $w(l)$은 위너 필터 계수이다. 따라서, 식 (2)의 비용 함수는 위너 필터 계수 $w(l)$에 관한 2차 함수이다. 최적 위너 필터 계수는 식 (2)를 $w(l)$로 미분한 후에 0으로 놓음으로써 구할 수 있다. 식 (4)에서 그 결과를 나타내었다.

$$\frac{\partial \zeta}{\partial w(m)} = 2E[e(n)x(n-m)] = 0, \ \forall m \tag{4}$$

식 (4)는 잘 알려진 직교 원칙(orthogonality principle)이다. 즉, 선형 필터 계수가 최적화되면 입력 신호와 추정 오차간의 내적은 0이 된다. 식 (4)를 좀 더 전개하면 식 (5)와 같이 자기 상관도 및 상호 상관도를 사용하여 표현이 가능하다.

$$\sum_{l=-\infty}^{\infty} w(l) r_x(m-l) = r_{dx}(m),\ \forall m \tag{5}$$

식 (5)에서 $r_x(m-l) = E[x(n-l)x(n-m)]$, $r_{dx}(m) = E[d(n)x(n-m)]$이다. 앞서 언급한 바와 같이 음성 신호와 잡음 신호가 통계적으로 독립이라면 $E[d(n-l)v(n-m)] = 0,\ \forall l,m$이므로 위의 자기 상관도 및 상호 상관도는 식 (6)과 식 (7)로 각각 정리될 수 있다.

$$\begin{aligned} r_x(m-l) &= E[x(n-l)x(n-m)] \\ &= E[(d(n-l)+v(n-l))(d(n-m)+v(n-m))] \\ &= E[d(n-l)d(n-m)] + E[v(n-l)v(n-m)] \\ &= r_d(m-l) + r_v(m-l) \end{aligned} \tag{6}$$

$$\begin{aligned} r_{dx}(m) &= E[d(n)(d(n-m)+v(n-m))] \\ &= r_d(m) \end{aligned} \tag{7}$$

식 (6)과 식 (7)을 식 (5)에 대입하면 식 (8)과 같이 정리된다.

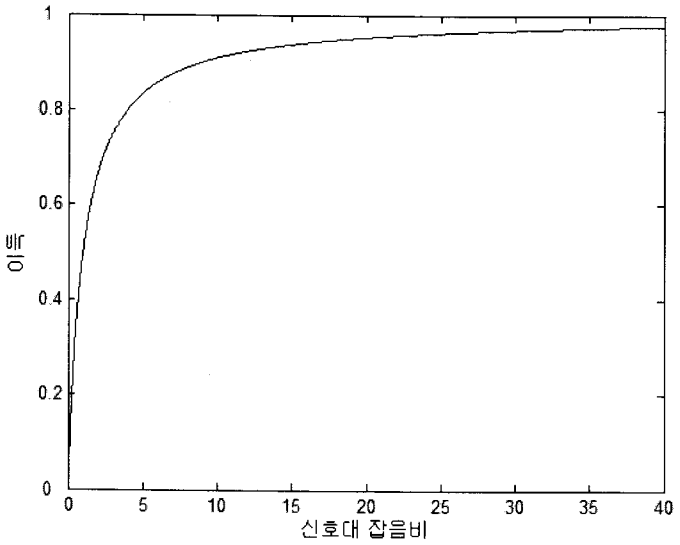
$$\sum_{l=-\infty}^{\infty} w(l)(r_x(m-l) + r_v(m-l)) = r_d(m),\ \forall m \tag{8}$$

주파수 영역에서 최적 위너 필터의 응답을 추정하기 위하여 식 (8)의 양변에 이산 시간 푸리에 변환(DTFT: discrete-time Fourier transform)을 취하면 식 (9)와 같이 표현되며 그것에 역 이산 시간 푸리에 변환(IDTFT: inverse DTFT)을 취하여 시간 영역의 최적 위너 필터 계수로 삼는다.

$$\begin{aligned} W(\omega) &= \frac{P_d(\omega)}{P_d(\omega) + P_v(\omega)} \\ &= \frac{SNR(\omega)}{1 + SNR(\omega)} \end{aligned} \tag{9}$$

식 (9)에서 $P_d(\omega), P_x(\omega)$는 각각 입력 음성신호 및 가산성 잡음의 전력 스펙트럼이다. ω는 디지털 주파수이다. 식 (9)에서 알 수 있듯이 위너 필터는 신호대 잡음비가 높은 주파수 대역은 그대로 유지시키고 신호대 잡음비가 낮은 주파수 대역은 감쇄시켜주는 역할을 한다. 신호대 잡음비에 따른 위너 필터의 특성은 <그림 1>과 같다.

음성 부호화기 및 음성 인식기에서의 적용을 위해서는 초기 입력 신호의 100~200 ms 구간을 순수 잡음 신호로 간주하고 잡음의 전력 스펙트럼을 먼저 추정한다. 일반적으로 정상성(stationary) 잡음의 경우 약 2초 동안은 그 특성이 크게



<그림 1> 신호대 잡음비에 따른 위너 필터 이득의 특성

변하지 않는다고 가정할 수 있으므로 고립단어 음성인식기의 전처리기에 응용될 때는 잡음 전력 스펙트럼의 재추정을 일반적으로 수행하지 않는다. 그렇지만, 음성 부호화기처럼 긴 구간 동안의 입력이 들어오는 경우에서는 에너지 기반의 음성 검출기가 실시간으로 동작하여 잡음 구간을 추정하게 되고 그에 따라서 잡음의 전력 스펙트럼이 갱신된다. 잡음의 전력 스펙트럼 추정치를 알고 있을 때, 음성의 전력 스펙트럼은 식 (10)과 같이 구할 수 있으며, 각 주파수 영역에서 신호대 잡음비를 구할 수 있으므로 식 (9) 및 IDTFT를 이용하여 시간 영역의 최적 위너 필터의 계수를 추정할 수 있다.

$$\widehat{P}_d(\omega) = P_x(\omega) - P_v^{(EST)}(\omega) \tag{10}$$

식 (10)에서 $\widehat{P}_d(\omega), P_x(\omega), P_v^{(EST)}(\omega)$는 각각 음성의 전력 스펙트럼 추정치, 잡음이 포함된 입력 신호의 전력 스펙트럼, 초기 입력신호 혹은 음성 검출기를 이용하여 추정한 잡음의 전력 스펙트럼이다.

위의 과정을 통한 표준 위너 필터는 시간 영역에서 무한 단위 응답(IIR: infinite impulse response) 형태를 갖게 되는데 이는 실제 시스템에 적용할 수 없다. 따라서, 일반적으로는 해밍창(Hamming window)을 이용하여 유한 단위 응답(FIR: finite impulse response) 형태로 바꾸어 잡음 제거를 수행한다.

2.2. 제안된 위너 필터

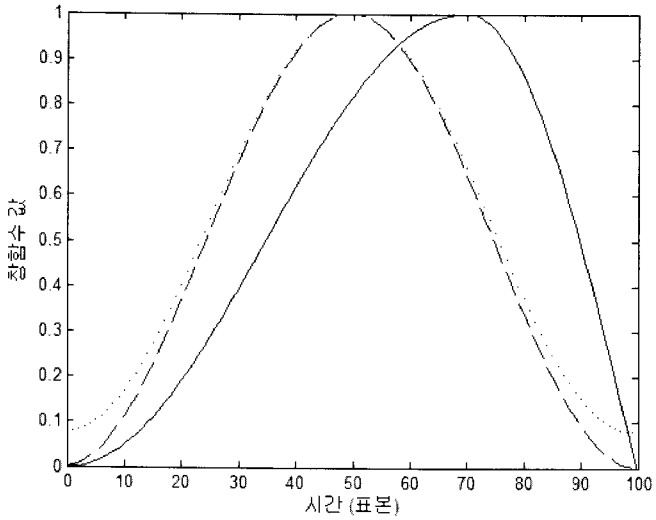
본 연구에서 위너 필터의 최적화 방식은 크게 두 가지로 요약이 가능하다. 첫 번째로는 비대칭 윈도우 사용에 따른 잡음 제거 알고리즘에서 소요되는 시간 지연의 최소화이며, 두 번째로는 신호대 잡음비의 변형을 이용한 잡음 제거 후의 음질 향상이다.

2.2.1. 비대칭 윈도우의 적용

일반적으로 잡음의 통계량이 정상성이라 하더라도 음성의 통계량은 잡음과 비교할 수 없을 정도의 비정상성을 띈다. 따라서, 위너 필터는 음성 신호가 매우 짧은 구간에서만 정상성의 주파수 스펙트럼을 가진다는 가정 하에서 약 10 ms 간격으로 추정된다. 위너 필터의 구현에 필수적인 주파수 분석은 단구간에서 수행될 경우에 해밍창, 해닝창(Hanning window) 등을 이용한 신호의 추출이 먼저 행해진다. 즉, 주파수 분석은 창 함수가 가장 높은 값을 갖는 입력 신호 표본 값에 대해서 수행되었다고 볼 수 있다. 여기서 주목할 것은 앞서 언급한 단구간 창 함수가 그것의 중앙에서 가장 큰 값을 가지므로 신호 버퍼링에 따른 시간 지연이 수반될 수 있다. 따라서, 본 연구에서는 단구간 주파수 분석에 사용될 창함수의 최대치가 가급적 오른쪽으로 치우친 형태가 되도록 만들어서 사용한다. 치우침의 정도는 전체적인 잡음 제거기의 성능을 어느 정도 유지하면서 최대가 되도록 한다. 본 연구에서는 식 (11)의 비대칭형 창함수를 사용하였다[13].

$$h(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{P_1}\right) & ,0\le n<n_0\\ \cos\left(\dfrac{2\pi(n-n_0)}{P_2}\right) & ,n_0\le n<N\end{cases} \tag{11}$$

식 (11)에서 P_1, P_2는 비대칭 창함수의 왼쪽 및 오른쪽 부분을 나타내기 위한 주기값이며, n_0 및 N은 최대치가 존재하는 위치 및 창함수 전체의 길이를 나타낸다. <그림 2>에서 대표적으로 사용되는 대칭형 창함수와 본 연구에서 사용한 비대칭형 창함수의 형태를 표시하였다. 만약, 신호 버퍼링의 크기가 80이라고 하면, <그림 2>의 예시에서 알 수 있듯이, 해닝 및 해밍 창함수를 사용할 경우에 미래의 입력을 요구하므로 신호 버퍼링에 의한 시간 지연이 수반될 수 있다. 반면, 비대칭 창함수의 경우는 과거의 입력만으로도 분석이 가능하므로 신호 버퍼링에 의한 시간 지연을 피할 수 있다.



<그림 2> 길이가 100인 대칭 및 비대칭 창함수의 형태 예시 (실선: 식 (10)에서 $P_1 = 139, P_2 = 119, n_0 = 70$일 때의 창함수, 점선: 해밍 창함수, 파선: 해닝 창함수)

2.2.2. 신호대 잡음비의 변형

식 (9)에서 나타낸 바와 같이 위너 필터의 이득은 주파수 영역에서 정의되는 입력 신호의 신호대 잡음비의 함수로서 정의된다. 그런데, 식 (9)에서 정의된 위너 필터의 이득은 식 (2)의 비용함수를 최소화하는 과정에서 구해진 것이다. 즉, 인간의 청각특성을 전혀 고려하지 않고 정의되었고 신호간의 차이만을 최소화할 수 있는 필터이다. 인간의 청각 특성은 여러 가지 비선형적인 요소를 포함하고 있기 때문에 그것을 모두 반영할 수 있는 수학적 모델링이나 비용함수를 정의하기는 매우 어렵다. 그렇지만, ETSI 표준에 채택되어 있는 바와 같이 입력 신호의 신호대 잡음비를 변형함에 의해서 음질 향상을 꾀할 수 있음이 알려져 있다[4]. 따라서 본 논문에서는 신호대 잡음비의 함수로 정의되는 위너 필터의 이득을 변형 시킬 수 있는 파라미터를 제시하고 그것을 인지적 측면에서 정의되는 비용함수를 최소화함에 의해 최적치를 구하는 방식을 제안한다. 식 (12)에서 변형된 위너 필터의 이득을 주파수 영역에서 표시하였다.

$$W_M(k) = \frac{SNR^{\alpha}(k)}{1 + SNR^{\alpha}(k)}, 0 < \alpha \le 1 \tag{12}$$

여기서, $W_M(k)$는 이산 푸리에 변환(DFT: discrete Fourier transform) 영역에서 주어지는 k번째 이산 주파수에서의 변형된 위너 필터 이득이다. ETSI 표준에서 제시

되는 위너 필터는 식 (12)에서 $\alpha = 0.5$로 두고 구한 것과 동일하다[4]. α가 커지면 신호대 잡음비가 조금만 높아져도 입력 신호를 그대로 보존 시키며, 신호대 잡음비가 낮은 부분은 심하게 값을 감쇄시키는 특성을 가진다. 반대로 α가 작아지면 신호대 잡음비가 낮은 영역에서 입력 신호의 감쇄를 적게 하고, 높은 영역에서는 감쇄가 상대적으로 높아지는 특성을 갖는다.

파라미터 α를 최적화하기 위한 비용함수는 식 (13)과 같다.

$$\begin{aligned} J &= \frac{1}{TL/2}\sum_{t=0}^{T-1}\sum_{k=1}^{L/2}\left(\log|D(t,k)| - \log|\hat{D}(t,k)|\right)^2 \\ &= \frac{1}{TL/2}\sum_{t=0}^{T-1}\sum_{k=1}^{L/2}\left(\log|D(t,k)| - \log|W_M(t,k)X(t,k)|\right)^2 \\ &= \frac{1}{TL/2}\sum_{t=0}^{T-1}\sum_{k=1}^{L/2}\left(\log|D(t,k)| - \log|X(t,k)| - \log W_M(t,k)\right)^2 \end{aligned} \tag{13}$$

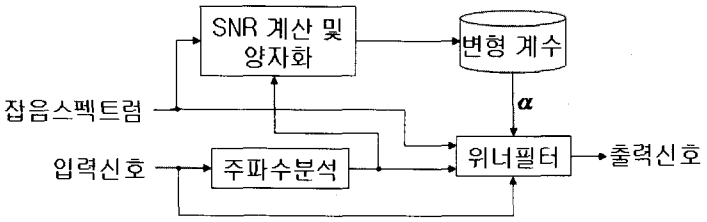
식 (13)에서 T는 최적화에 사용된 총 음성 신호의 분석 프레임 수, L은 DFT의 크기, t, k는 각각 분석 프레임 및 이산 주파수의 인덱스, $D(t,k)$는 잡음이 섞이지 않은 훈련용 음성의 DFT, $X(t,k)$는 잡음이 섞인 훈련용 음성의 DFT, $\hat{D}(t,k)$는 잡음이 제거된 음성신호이다. 인간의 청각은 신호의 세기를 로그 스케일로 느끼므로 식 (13)은 인지적 비용함수로 볼 수 있다. 식 (13)은 α에 관한 함수이긴 하지만, 비선형적인 요소를 포함하고 있기 때문에 폐형 해(closed form solution)를 구하기 어렵다. 따라서, 본 연구에서는 식 (14)와 같이 적당한 초기치로부터 시작하여 비용함수의 미분을 통해 얻어지는 그레디언트(gradient)를 이용한 최적화 기법을 사용한다.

$$\alpha(n+1) = \alpha(n) - \mu\nabla_{\alpha(n)}J \tag{14}$$

μ는 학습 (learning rate), ∇는 그레디언트 연산을 나타낸다.

식 (14)를 직접적으로 적용하여도 잡음 제거 후의 음질을 더 높일 수 있는 변형된 위너 필터가 구해질 수 있다. 그렇지만, 더욱 높은 성능 향상을 얻기 위해서 식 (13)의 비용 함수를 입력신호의 신호대 잡음비에 따라서 좀 더 세분화하여 나누고 α 및 식 (14)에서 제시된 학습 알고리즘 역시 세분화된 신호대 잡음비에 최적화되도록 나눈다. 이들을 식 (15)와 식 (16)에 자세히 나타내었다.

$$J_q = \frac{1}{Q}\sum_t\sum_k\left(\log|D(t,k)| - \log|X(t,k)| - \log W_M(t,k)\right)^2, \quad SNR_q = Quant(SNR(t,k))\text{를 만족하는 } \forall\,(t,k) \tag{15}$$



<그림 3> 제안된 방식의 위너 필터를 이용한 잡음 제거

$$\alpha_q(n+1) = \alpha_q(n) - \mu \nabla_{\alpha_q(n)} J_q \tag{16}$$

$$\nabla_{\alpha_q(n)} = -\frac{2}{Q}\sum_t\sum_k (\log|D(t,k)| - \log|X(t,k)| - \log W_M(t,k)) \frac{\log SNR(t,k)}{1+SNR^{\alpha(n)}(t,k)}, \quad SNR_q = Quant(SNR(t,k)) \text{를 만족하는 } \forall\ (t,k) \tag{17}$$

식 (15), (16), (17)에서 $Quant(\bullet)$는 양자화 연산을 나타내며 코드북 인덱스를 출력한다. Q는 시간-주파수 영역에서 $SNR_q = Quant(SNR(t,k))$인 총 이산 스펙트럼의 개수이다. α_q는 q번째 신호대 잡음비 코드북을 갖는 비용함수에 사용된 파라미터이다. 식 (16)에 의한 α_q의 갱신을 위해서 식 (17)은 전체 훈련 데이터베이스에서 구해진다. 즉, 식 (16)의 훈련은 소위 'train-by-epoch'에 의해서 이루어진다.

위의 과정들을 모두 반영한 제안된 위너 필터 기법을 <그림 3>에 나타내었다. 그림에서 위너 필터의 설계에 필요한 음성 신호의 스펙트럼 추정은 음성 검출기로부터 검출된 비음성 구간에서 잡음의 스펙트럼을 정밀하게 추정하고 입력 신호의 스펙트럼으로부터 그것을 빼줌으로써 가능하다. 그 후, 추정된 음성 스펙트럼과 잡음 스펙트럼의 비 및 그것의 양자화를 통해서 최적의 변형 계수 α가 출력되고 위너 필터가 구현되며 최종적으로 잡음이 줄어든 음성신호를 얻게 된다. 주파수 영역에서의 위너 필터 추정은 입력 신호대 잡음비의 양자화만 수행되면 테이블 검색으로 쉽게 얻어지므로 변형계수 도입에 따른 부가적인 계산량 증가는 무시할 수 있다.

3. 실험 및 결과

성능의 지표로서 PESQ를 사용하여 객관적 음질 평가를 수행하였다. PESQ 스코어는 -0.5 ~ 4.5 사이의 값을 출력하며 주관적 음질 평가의 지표인 MOS (mean opinion score)와 최대한 유사한 값을 갖도록 설계되어 있다. 알고리즘 별로 혹은

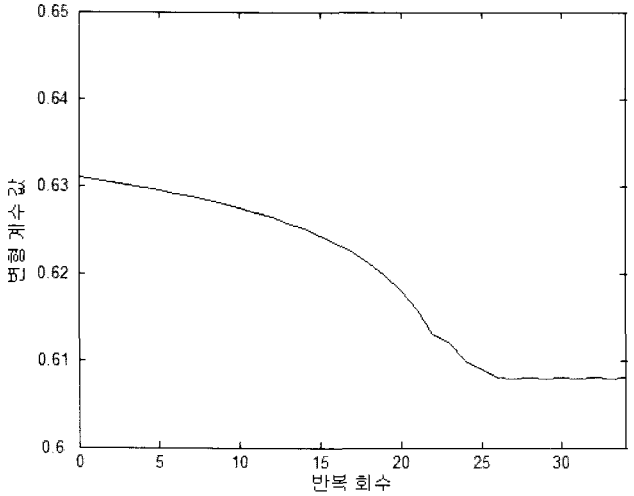
잡음의 종류 별로 잡음 제거 전후의 음질 향상 정도를 분석하였다.

남녀 각 2명이 발성한 100 문장이 테스트로 사용되고, 문장의 길이는 대부분 10 초 내외이다. 신호대 잡음비의 정량화 및 최적 변형 계수 α_q를 구하기 위한 훈련과정에 40 문장이 사용되고 인위적으로 다양한 성격의 잡음을 섞었으며 문장 단위의 신호대 잡음비는 5 dB로 맞추었다. 음질 평가를 위하여 테스트 문장에 인공적으로 섞은 잡음의 종류는 백색잡음, 사무실 잡음, 배블성 잡음, 자동차 주행 잡음 등이며, 모든 신호는 8 kHz 표본화율, 16 bit 양자화로 녹음되었다. 자동차 잡음은 70 km/h로 주행 중인 차량에서 수집하였다. 주행 상황에서 창문은 모두 닫았고, 오디오 시스템은 끈 상태에서, 에어컨은 적당히 킨 상태에서 수집하였다. 수집한 잡음은 신호대 잡음비가 35 dB 이상인 테스트 문장에 가산적으로 섞이며, 성능 변이를 분석하기 위해서 신호대 잡음비가 0 dB, 5 dB, 10 dB, 15 dB, 20 dB가 되도록 인공적으로 맞추었다.

제안된 위너 필터는 매 10 ms 단위로 추정되며, 에너지 기반의 음성 검출기를 이용하여 비음성구간을 추정한다. 주파수 분석을 위한 단구간 프레임의 크기는 100 샘플이며, 비대칭 창함수의 모양은 식 (11) 및 <그림 2>에서 제시한 것과 동일하다. 신호 버퍼링에 의한 지연을 막기 위해서 현재 입력 버퍼 80 샘플과 과거의 버퍼 20 샘플이 단구간 분석 프레임에 이용된다. 주파수 분석을 위하여 128 FFT (fast Fourier transform)를 사용하였으며 길이 61의 위너 필터가 추정된다. 신호대 잡음비의 양자화를 위해서 LBG (Linde-Buso-Gray) 알고리즘이 사용되었으며 총 코드북의 수는 64 개였다. 양자화된 신호대 잡음비의 최소치는 0 dB, 최대치는 40 dB로 두었다. 변형 계수를 학습하기 위한 학습률은 10^{-4}이었다. 식 (16)에서 변형 계수 α_q의 초기치는 0.5 ~ 1.0 사이에서 무작위로 선택되었으며 대부분의 경우에서 충분히 수렴하였을 때, 0.6 ~ 0.7 사이의 값을 가졌다. 양자화된 신호대 잡음비가 5 dB일 때의 학습곡선의 예를 <그림 4>에 나타내었다. 식 (9)에 나타낸 표준 위너 필터는 $\alpha_q = 1.0$인 것과 등가이다. 실험 결과에 따르면 입력 신호대 잡음비가 5 dB일 때, 최소 자승 오차(MMSE: minimum mean-squared error) 기반의 표준 위너 필터는 입력 신호의 전력 스펙트럼을 2.38 dB, 제안된 인지적 오차 기반의 위너 필터는 3.49 dB 감쇄시킨다. 잡음 제거를 위한 입력 신호의 감쇄량은 전반적으로 제안된 위너 필터에서 더 크다.

실제 VoIP 상황과 최대한 일치시키기 위해서 잡음 제거 후의 신호를 G.711, G.723.1, G.729A 등의 음성 부호화기 및 복호화기를 통과시킨 신호에 대해서 음질 측정을 수행하였다. 제안한 알고리즘의 성능 비교를 위하여 IS-127 EVRC에 내장되어있는 잡음 제거기 및 ETSI 표준안에 제안되어 있는 분산형 음성인식환경 시스템의 전처리기용 잡음제거기와의 PESQ 스코어를 비교하였다.

<표 1>, <표 2>, <표 3>은 G.711, G.723.1, G.729A에 대한 각각의 평균 PESQ 스코어를 나타낸 것이다. 대부분의 잡음 조건에서 제안된 방식이 다른 방식들에



<그림 4> 신호대 잡음비가 5 dB일 때, 변형 계수 α_q의 학습 곡선

비해서 높은 스코어를 가짐을 알 수 있다. 신호대 잡음비가 낮아짐에 따라서 제안한 알고리즘에 의한 성능 향상의 폭이 더욱 커졌다. G.723.1 및 G.729A의 결과에서 배블성 잡음일 경우에 ETSI 잡음제거기의 성능이 오히려 낮아짐을 알 수 있는데, 부호화 복호화 과정에서의 신호 왜곡 요소가 제안한 알고리즘의 결과에 더 큰 영향을 주었기 때문으로 판단할 수 있다. 전반적인 성능의 순위는 제안된 방식, ETSI 용 잡음 제거기, EVRC 용 잡음 제거기의 순서로 볼 수 있다. 제안된 방식의 잡음 제거 전후의 평균 PESQ 스코어 이득을 보면, G.711일 경우 0.44, G.723.1일 경우 0.30, G.729A일 경우 0.28 정도였다. 특히, 차량 잡음일 경우에 다른 잡음 제거기에 비해서 우수한 성능을 나타내었다. 시간 지연 측면에서 EVRC는 유리할 수 있지만 음질 측면에서 가장 나쁜 성능을 보였다.

<표 1> G.711의 부호화/복호화기를 통과하였을 때의 PESQ 스코어

	백색 잡음					사무실 잡음				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.52	2.20	1.89	1.61	1.42	2.77	2.44	2.11	1.76	1.44
EVRC	3.07	2.73	2.32	1.83	1.39	3.13	2.78	2.42	2.02	1.64
ETSI	3.17	2.84	2.48	2.09	1.69	3.18	2.80	2.44	2.04	1.64
Proposed	3.27	2.96	2.64	2.27	1.82	3.25	2.88	2.52	2.13	1.74
	배블성 잡음					차량 잡음				
None	2.73	2.40	2.11	1.79	1.51	3.61	3.28	2.97	2.64	2.28
EVRC	3.11	2.76	2.45	2.04	1.67	3.84	3.52	3.22	2.90	2.57
ETSI	3.14	2.76	2.48	2.11	1.73	3.85	3.60	3.26	2.92	2.55
Proposed	3.18	2.81	2.49	2.13	1.75	3.89	3.65	3.30	3.01	2.63

<표 2> G.723.1 (6.3 kbps)의 부호화/복호화기를 통과하였을 때의 PESQ 스코어

	백색 잡음					사무실 잡음				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.64	2.34	2.03	1.70	1.44	2.87	2.58	2.28	1.91	1.54
EVRC	2.95	2.70	2.36	1.90	1.40	3.05	2.77	2.46	2.07	1.68
ETSI	3.07	2.85	2.55	2.20	1.78	3.09	2.82	2.51	2.12	1.72
Proposed	3.10	2.85	2.59	2.27	1.86	3.12	2.84	2.52	2.15	1.78
	배블성 잡음					차량 잡음				
None	2.84	2.55	2.27	1.92	1.59	3.29	3.09	2.84	2.55	2.22
EVRC	3.04	2.77	2.48	2.11	1.72	3.40	3.21	2.97	2.68	2.36
ETSI	3.07	2.79	2.52	2.19	1.79	3.40	3.26	3.02	2.75	2.41
Proposed	3.06	2.80	2.47	2.15	1.76	3.45	3.35	3.12	2.90	2.54

<표 3> G.729A의 부호화/복호화기를 통과하였을 때의 PESQ 스코어

	백색 잡음					사무실 잡음				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.71	2.40	2.06	1.71	1.43	2.93	2.63	2.31	1.92	1.51
EVRC	2.98	2.74	2.40	1.93	1.40	3.09	2.80	2.48	2.08	1.67
ETSI	3.13	2.89	2.59	2.23	1.79	3.13	2.84	2.51	2.09	1.67
Proposed	3.14	2.86	2.60	2.27	1.85	3.15	2.85	2.52	2.11	1.72
	배블성 잡음					차량 잡음				
None	2.90	2.60	2.30	1.94	1.59	3.36	3.15	2.87	2.57	2.23
EVRC	3.08	2.80	2.51	2.13	1.74	3.49	3.32	3.06	2.74	2.40
ETSI	3.10	2.83	2.54	2.19	1.77	3.46	3.30	3.03	2.73	2.37
Proposed	3.11	2.83	2.47	2.14	1.75	3.52	3.40	3.15	2.89	2.51

<표 4> 잡음제거 알고리즘별 지연시간 및 10 ms 처리에 필요한 시간

	EVRC	ETSI	Proposed
지연시간 (ms)	0	40	3.75
10ms당 처리시간 (ms)	0.016	0.187	0.139

VoIP 환경에서는 음질뿐만 아니라 시간 지연도 중요한 요소이기 때문에 <표 4>에 잡음제거 알고리즘별 신호 지연양을 나타내었으며 부가적으로 10 ms 처리를 위해 필요한 시간을 나타내었다. 측정 환경은 Pentium-4 3.2 GHz 및 1 GByte RAM을 장착한 PC 이다. 제안된 알고리즘의 신호 지연 요소는 비인과성 위너 필터의 설계에 따른 것이므로 위너 필터 길이의 절반만큼의 신호지연은 피할 수 없다. 잡음제거 알고리즘에 의한 지연시간의 양은 ETSI 잡음제거, 제안된 방식, EVRC 잡음제거의 순이었다. ETSI 잡음제거의 경우 위너 필터에 의한 잡음제거가 두 단계에 걸쳐서 수행되고 신호 분석을 위한 FFT의 길이도 256이므로 40 ms의 긴 시간 지연을 요구한다. 이는 ITU-T G.114의 권고안의 최대 시간 지연의 약 1/4에 이르

므로 바람직하지 못하다고 하겠다. 반면, 제안한 방식은 부담되지 않을 정도의 시간 지연을 일으키지만 가장 좋은 음질 향상을 얻을 수 있었다. 처리시간 측면에서는 EVRC가 압도적으로 적게 들지만 비교 대상의 알고리즘들 역시 10 ms 보다 훨씬 적은 시간을 요구하므로 큰 고려대상은 아닌 것으로 판단된다.

4. 결 론

본 논문에서는 가산적인 잡음을 제거하기 위해서 위너 필터의 최적화 방식에 대해서 연구하였다. 위너 필터의 최적화는 변형된 신호대 잡음비의 제안 및 그에 사용된 파라미터의 최적화를 인지적 요소의 비용함수를 최소화함으로써 수행되었다. VoIP 시스템의 잡음제거기로 사용되기 위하여 알고리즘에서 소요되는 시간 지연 요소 역시 고려되었으며 이를 위해서 비대칭형 창함수를 이용한 주파수 분석을 수행하였다. 정확한 성능의 측정을 위하여 본 연구에서 구현한 잡음 제거기는 실제 VoIP 용 음성 부호화기의 전처리기로 사용되었으며 PESQ 스코어를 측정한 결과 EVRC 및 ETSI 잡음 제거기에 비해서 우수한 성능을 보였다. 따라서, 본 연구에서 제안한 잡음 제거 알고리즘은 VoIP 분야에서 음성 신호에 섞이는 가산적 잡음을 제거하는 데에 효과적으로 사용될 수 있음을 알 수 있었다.

향후 연구로서 좀 더 정확한 인지적 특성을 반영하기 위해서 바크 (Bark) 혹은 멜 (mel) 스케일 주파수 영역에서 스펙트럼 왜곡 비용함수를 사용할 수 있겠고, 비용함수 역시 각 주파수에 성분에 대하여 각각 정의하여 최적화할 수 있을 것이다. 또한, 잡음 제거 후의 음성 인식률 향상의 측정 역시 향후 연구에 포함될 수 있다.

참 고 문 헌

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113-120, 1979.

[2] Enhanced Variable Rate Codec (EVRC), *Speech service option 3 for wideband spectrum digital systems*, 1996.

[3] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 32, No. 6, pp. 1109-1121, 1984.

[4] ETSI ES 202 212, *Speech processing, transmission and quality aspects(STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm;*

compression algorithms; back-end speech reconstruction algorithm, pp. 1-21, 2005.

[5] S. Gannot, D. Burshtein, E. Weinstin, "Iterative and sequential Kalman filter based speech enhancement algorithms", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 4, pp. 373-385, 1998.

[6] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, 1996.

[7] ITU-T Recommendation P. 862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codec*, 2001.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 5, pp. 504-512, 2001.

[9] R. J. McAulay, M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 2, pp. 137-145, 1980.

[10] K. K. Paliwal, A. Basu, "A speech enhancement method based on Kalman filtering", *Proc. ICASSP*, pp. 177-180, 1987.

[11] B. Widrow, S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.

[12] J. Turunen, P. Loula, T. Lipping, "Assessment of objective voice quality over best-effort networks", *Computer Communications*, Vol. 28, No. 5, pp. 582-588, 2005.

[13] ETSI ES 300 726, *Digital cellular telecommunications system; Enhanced full rate (EFR) speech transcoding (GSC06.60)*, p. 20, 1997.

[14] 박정식, 오영환, "MMSE estimator 기반의 적응 콤 필터링을 이용한 잡음 제거", *말소리*, 제60호, pp. 181-190, 2006.

접수일자: 2007년 11월 10일
게재결정: 2007년 12월 16일

▶ 정상배(Sangbae Jeong) : 교신저자
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6285
E-mail: sangbae@icu.ac.kr

▶ 이성독(Sungdoke Lee)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6285
E-mail: sdlee@icu.ac.kr

▶ 한민수(Minsoo Hahn)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6123
E-mail: mshahn@icu.ac.kr