말소리와 음성과학 제3권 제1호 (2011)
pp. 87~94

위상 정보를 고려한 로그멜 영역에서의 2단계 선험 SNR 추정

Two-step *a priori* SNR Estimation in the Log-mel Domain Considering Phase Information

이 윤 경[1] · 권 오 욱[2]
Lee, Yun-Kyung* · Kwon, Oh-Wook**

ABSTRACT

The decision directed (DD) approach is widely used to determine a priori SNR from noisy speech signals. In conventional speech enhancement systems with a DD approach, a priori SNR is estimated by using only the magnitude components and consequently follows a posteriori SNR with one frame delay. We propose a phase-dependent two-step a priori SNR estimator based on the minimum mean square error (MMSE) in the log-mel spectral domain so that we can consider both magnitude and phase information, and it can overcome the performance degradation caused by one frame delay. From the experimental results, the proposed estimator is shown to improve the output SNR of enhanced speech signals by 2.3 dB compared to the conventional DD approach-based system.

Keywords: phase modeling, speech enhancement, speech separation, MMSE, decision-directed, a priori SNR

1. 서론

잡음 제거 기술은 실제 환경에서 음성 인식 시스템을 적용하기 위하여 필수적이다. 실제 환경에서의 음성 신호는 대부분 다양한 종류의 배경 잡음을 포함하며, 이러한 배경 잡음은 음성 인식 시스템의 성능을 저하시키는 주요한 원인이 된다. 단일 채널 음성 인식 시스템에서 잡음 요인을 제거하고 음성 신호를 향상시키는 기술은 음질을 높일 뿐 아니라 음성 인식률을 개선하고 음성의 명료도를 높여 피로감을 감소시키는 효과가 있다.

원하는 음성 신호를 추정하고 음성을 향상시키기 위해 널리 사용되는 방법은 선험(*a priori*) 신호대잡음비(signal-to-noise ratio: SNR)를 추정한 후 이를 다시 혼합 신호에 곱함으로써 배경잡음을 제거하는 기술이다. Decision directed(DD) 접근법은 선험 SNR을 결정하기 위한 대표적인 기술로, 추출된 음성신호의 잔여 잡음에 의해 발생되는 음악적 잡음(musical noise)을 감소시키며 성능이 우수하다고 알려지고 있다[1]-[3]. 하지만 기존의 DD 접근법을 비롯한 음성 향상 기술은 선험 SNR을 계산하고 음성을 향상시키는 과정에서 음성의 크기 성분(magnitude component)만을 이용한다. 따라서 음성의 위상 성분이 음성 신호에 대한 정보를 가지며 위상 정보를 이용한 모델이 사람의 음성 인지(human speech perception)와 음성인식에 유용하다는 최근의 연구결과[4]-[6]를 충분히 반영하지 못하고 있다. 또한 기존의 DD 접근법에서는 현재의 선험 SNR을 계산하기 위해 이전 프레임에서 추정된 음성의 스펙트럼을 이용하기 때문에 선험 SNR은 한 프레임의 시간 지연을 가지고 후험(*a posteriori*) SNR의 형태를 따라간다는 단점이 있다.

이러한 기존의 문제점들을 해결하기 위해, 본 논문에서는 위상 구간을 0으로 만들지 않고 음성 신호의 크기와 위상 성분을 모두 고려하기 위해 혼합 신호의 전력 스펙트럼 벡터를 로그멜 스펙트럼 벡터로 변환시킨 후 선험 SNR을 추정하고 음성 신호를 향상시킴으로써 음성 향상의 성능을 높인다. 또한, 추정된 선험 SNR을 MMSE 기반의 선험 SNR 추정법을 이용하여 재추정함으로써 DD 접근법의 장점은 유지하면서 한 프레임의 시간 지연을 갖는 선험 SNR의 문제를 해결하도록 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 신호 모델링에

1) 충북대학교 제어로봇공학과 yklee@cbnu.ac.kr
2) 충북대학교 제어로봇공학과 owkwon@cbnu.ac.kr, 교신저자
이 논문은 2010년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었습니다.

접수일자: 2010년 12월 31일
수정일자: 2011년 3월 18일
게재결정: 2011년 3월 25일

대하여 소개한다. 제3장에서는 기존의 DD 접근법을 간략히 소개하고, 로그멜 영역에서 위상 정보를 고려한 선험 SNR 추정 기술과 위상 기반의 DD 접근법을 설명한 후 MMSE 기반의 이단계 선험 SNR 추정 기술과 이를 이용한 위상 기반의 음성 향상 기술을 설명한다. 제4장에서는 음성의 크기와 위상 성분을 이용하여 음성 신호 향상을 수행한 실험 결과를 제시하고, 마지막으로 제5장에서는 결론을 맺는다.

2. 신호 모델링

단일 마이크로폰을 이용해 얻어진 원 음성 신호와 잡음 신호를 각각 $x(t)$, $n(t)$라고 할 때, 혼합 음성 신호 $y(t)$는 두 신호의 간단한 합으로 얻어지며 다음과 같이 정의된다.

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

$x(t)$와 $n(t)$의 크기 성분(magnitude spectrum)을 각각 $|X|$, $|N|$이라고 할 때, 혼합 신호의 스펙트럼 크기 $|Y|$는 다음과 같이 정의된다.

$$|Y|^2 = |X|^2 + |N|^2 + 2\cos(\theta)|X||N|. \tag{2}$$

여기서, θ는 $|X|$와 $|N|$ 사이에 관련된 위상 차이이다. 일반적으로, 위상 구간 $2\cos(\theta)|X||N|$은 평균적으로 0이라고 가정하여 무시한다.

$$|Y|^2 = |X|^2 + |N|^2. \tag{3}$$

그러나 식 (1)을 로그멜 영역으로 변환하면 로그 함수에 의해 비선형 변환이 적용되고, 이로 인해 $|X|$와 $|N|$ 사이의 위상 구간은 더 이상 평균적으로 0이 되지 않는다. $f(x)$가 로그 함수라고 할 때, 비선형으로 변환된 pdf의 평균 $E_{p(f(x))}[f(x)]$는 원래 pdf의 평균을 비선형 변환한 $f(E_{p(x)}[x])$와 같지 않게 되기 때문이다[4].

멜 필터뱅크는 필터의 폭을 멜 스케일로 조정한 것으로, 저주파에 민감하고 고주파에서는 상대적으로 둔감한 달팽이관을 모델링하여 사람의 청각 특징에 맞추어 필터처리를 수행한다. 그림 1에 23차의 정규화한 멜 필터의 예를 나타내었다. 멜 필터뱅크 행렬 W에 의해 변환된 혼합 신호와 음성신호, 그리고 잡음 신호의 전력 스펙트럼 벡터($|Y|^2$, $|X|^2$, $|N|^2$)을 각각 $\tilde{Y}$, $\tilde{X}$, $\tilde{N}$라고 할 때, 식 (1)은 다음과 같이 변환된다.

$$\tilde{Y} = \tilde{X} + \tilde{N} + 2\sqrt{\tilde{X}\tilde{N}}\cos(\theta_{\tilde{X}} - \theta_{\tilde{N}}) \tag{4}$$

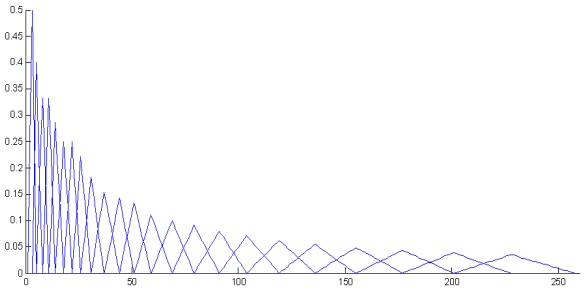

그림 1. 정규화한 멜 필터의 예 (차수: 23차)
Figure 1. An example of normalized mel filters (order: 23)

여기서, $\theta_{\tilde{X}}$와 $\theta_{\tilde{N}}$은 각각 $\tilde{X}$와 $\tilde{N}$의 위상 스펙트럼이며, 식 (4)는 $\tilde{X}$에 대해 두 근을 갖는다.

$$\tilde{X} = \left(-c_{\tilde{X}\tilde{N}}\sqrt{\tilde{N}} \pm \sqrt{(c_{\tilde{X}\tilde{N}}^2 - 1)\tilde{N} + \tilde{Y}}\right)^2, \tag{5}$$
$$c_{\tilde{X}\tilde{N}} = \cos(\theta_{\tilde{X}} - \theta_{\tilde{N}}).$$

3. 위상 기반 음성 향상 알고리듬

3.1 DD 접근법

혼합 신호로부터 선험 SNR을 결정하기 위해 널리 사용되는 방법은 DD 접근법이다[2, pp.219-231]. DD 접근법은 순간(instantaneous) SNR과 선험 SNR의 선형 결합으로 계산된다. 잡음 억제를 위한 파라미터로 사용되는 후험 SNR은 잡음과 혼합 신호의 전력 스펙트럼의 비로 정의되며, m번째 프레임, k번째 주파수 빈에서의 후험 SNR $\gamma(m,k)$는 다음과 같다.

$$\gamma(m,k) = \frac{|Y(m,k)|^2}{E\{|N(m,k)|^2\}} \tag{6}$$

잡음 신호의 전력 스펙트럼은 음성 부재인 구간 동안 추정된다. 선험 SNR은 잡음 신호와 음성 신호의 전력 스펙트럼의 비로 정의되며 다음과 같다.

$$\xi(m,k) = \frac{E\{|X(m,k)|^2\}}{E\{|N(m,k)|^2\}}. \tag{7}$$

순간 SNR은 다음과 같이 정의된다.

$$\vartheta(m,k) = \frac{|Y(m,k)|^2}{E\{|N(m,k)|^2\}} - 1. \tag{8}$$

식 (7)과 (8)의 선형 결합으로부터, DD 접근법은 가중치 파라미터 $0 < \alpha < 1$와 함께 다음과 같이 계산된다.

$$\xi(m,k)=E\left\{\alpha\frac{|X(m,k)|^2}{E\{|N(m,k)|^2\}}+(1-\alpha)\vartheta(m,k)\right\}. \quad (9)$$

위 식의 표현은 실제에서는 계산할 수 없기 때문에, 재귀적으로 근사하여 선험 SNR을 결정하는 데 사용한다.

$$\hat{\xi}(m,k)=\alpha\frac{|\hat{X}(m-1,k)|^2}{E\{|N(m-1,k)|^2\}}+(1-\alpha)\max[\gamma(m,k)-1,0]. \quad (10)$$

여기서, $|\hat{X}(m-1,k)|^2$은 이전 프레임에서 추정된 음성의 전력 스펙트럼 값이다.

3.2 로그멜 영역에서의 선험 SNR 추정

기존의 음성 향상을 위한 연구들은 일반적으로 위상을 0이라고 가정하여 무시하고, 음성의 크기 정보만을 이용한다. 위상 정보 역시 음성 신호에 대한 정보를 가지며, 잡음을 효과적으로 제거하고 음성 향상의 성능을 높이기 위해서는 음성의 크기 정보와 위상 정보를 함께 고려해 주어야 한다.

본 논문에서는 위상 성분이 0이 되지 않도록 하기 위해 혼합 음성 신호의 전력 스펙트럼 벡터를 로그멜 스펙트럼 벡터로 변환하여 로그멜 영역에서 선험 SNR을 추정하고, 이를 이용하여 음성 신호를 향상시킴으로써 음성 신호를 향상하는 과정에서 음성의 크기와 위상 정보를 모두 고려한다. 식 (5)의 ± 기호에 의한 둘 중 어떠한 부호를 사용해야 하는지 결정하는 것이 간단하지 않은 애매성을 해결하기 위해, 그림 2에 나타낸 원 음성신호와 잡음신호, 그리고 혼합 음성신호의 멜 스펙트럼 벡터의 관계에 코사인 법칙을 적용하여 $\widetilde{X}$와 이에 관련된 위상 구간을 대수적으로 계산한다[2, pp.97-110].

$$\widetilde{X}=\widetilde{Y}+\widetilde{N}-2c_{\widetilde{Y}\widetilde{N}}\sqrt{\widetilde{Y}\widetilde{N}}, \quad (11)$$

$$c_{\widetilde{Y}\widetilde{X}}=\cos(\theta_{\widetilde{Y}}-\theta_{\widetilde{N}}).$$

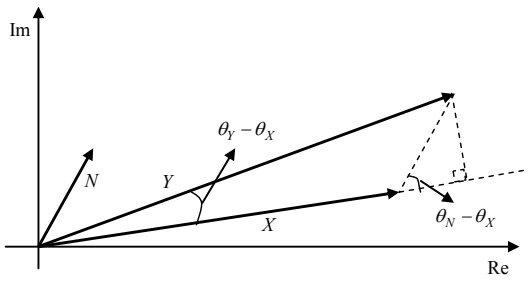

그림 2. 원 음성 신호와 잡음 신호, 그리고 혼합 음성 신호의 삼각관계

Figure 2. Diagram illustrating the trigonometric relationship of the clean, noise, and noisy signals.

여기서, m번째 프레임, k번째 주파수 빈에서의 위상 구간 $c_{\widetilde{Y}\widetilde{X}}(m,k)$는 코사인 법칙과 이전 프레임에서 추정된 스펙트럼 $\hat{\widetilde{X}}(m-1,k)$을 이용하여 $c_{\widetilde{X}\widetilde{N}}$과 관련된 재귀식으로 계산된다.

$$\sqrt{\widetilde{X}}=\frac{\left(\sqrt{\widetilde{Y}}\hat{c}_{\widetilde{Y}\widetilde{N}}-\sqrt{\widetilde{N}}\right)}{c_{\widetilde{X}\widetilde{N}}}, \quad (12)$$

$$\hat{c}_{\widetilde{Y}\widetilde{N}}(m,k)=\left(\frac{\sqrt{\hat{\widetilde{X}}(m-1,k)}}{\sqrt{\widetilde{Y}}(m-1,k)}c_{\widetilde{X}\widetilde{N}}(m-1,k)+\frac{\sqrt{\widetilde{N}(m,k)}}{\sqrt{\widetilde{Y}}(m,k)}\right).$$

음성 부재의 구간 동안 추정된 잡음 신호의 로그멜 전력 스펙트럼을 $\lambda_{\log(\widetilde{N})}(m,k)$라고 할 때, 음성의 위상 정보를 고려한 선험 SNR $\tilde{\xi}(m,k)$은 다음과 같이 정의된다.

$$\tilde{\xi}(m,k)=\frac{E\left\{\log(\widetilde{X}(m,k))\right\}}{\lambda_{\log(\widetilde{N})}(m,k)}, \quad (13)$$

$$\lambda_{\log(\widetilde{N})}(m,k)=E\left\{\log(\widetilde{N}(m,k))\right\}.$$

위상 기반의 순간 SNR $\tilde{\vartheta}(m,k)$는 다음과 같이 정의된다.

$$\tilde{\vartheta}(m,k)=\frac{E\left\{\log\left[\widetilde{Y}(m,k)+\widetilde{N}(m,k)-2\hat{c}_{\widetilde{Y}\widetilde{N}}\sqrt{\widetilde{Y}(m,k)\widetilde{N}(m,k)}\right]\right\}}{\lambda_{\log(\widetilde{N})}(m,k)}. \quad (14)$$

3.3 로그멜 영역에서의 DD 접근법

혼합 음성신호로부터 음성의 크기와 위상을 모두 고려한 선험 SNR을 결정하기 위하여, 기존의 DD 접근법을 차용하여 식 (13)과 (14)를 사용하여 계산된 선험 SNR과 순간 SNR의 선형 결합으로 정의된다.

$$\tilde{\xi}(m,k)=E\left\{\alpha\frac{\log(\widetilde{X}(m,k))}{\lambda_{\log(\widetilde{N})}(m,k)}+(1-\alpha)\bar{\xi}(m,k)\right\}, \quad (15)$$

$$\bar{\xi}(m,k)=\frac{\log\left[\widetilde{Y}(m,k)+\widetilde{N}(m,k)-2\widehat{c_{\widetilde{Y}\widetilde{N}}}\sqrt{\widetilde{Y}(m,k)\widetilde{N}(m,k)}\right]}{\lambda_{\log(\widetilde{N})}(m,k)}.$$

위 식을 재귀적으로 근사화하면 다음과 같다.

$$\hat{\tilde{\xi}}(m,k)=\alpha\left\{\frac{\log(\hat{\widetilde{X}}(m-1,k))}{\lambda_{\log(\widetilde{N})}(m-1,k)}\right\}+(1-\alpha)\max[\bar{\xi}(m,k),0]. \quad (16)$$

여기서, $0<\alpha<1$는 가중치이고, $\hat{\widetilde{X}}(m-1,k)$는 이전 프레임에서 추정된 로그멜 스펙트럼 값이다. 본 논문에서는 가중치 α를 0.98로 정하여 사용하였다.

3.4 이단계 선험 SNR 추정기

DD 접근법을 이용하여 결정된 선험 SNR은 이전 프레임에서 추정된 음성의 스펙트럼에 의존적인 경향을 보인다. 현재의 선험 SNR을 계산하기 위해 이전 프레임의 추정 결과를 이

용하기 때문에 하나의 시간 프레임 지연에서 후험 SNR의 형태를 따라가고, 이로 인해 음성 향상 성능을 감소시키는 원인이 된다.

DD 접근법의 장점을 유지하면서 이러한 문제를 해결하기 위해 본 논문에서는 최소평균자승오류(minimum mean square error: MMSE)에 기반을 둔 이단계 선험 SNR 추정 방법을 이용하여 DD 접근법을 통해 추정된 선험 SNR을 다시 한 번 재추정, 재정비하도록 한다[3].

3.4.1 MMSE 기반 선험 SNR 추정기

전력 스펙트럼 밀도 X^2에 대한 최소평균자승오류 추정은 다음과 같이 조건부 기댓값으로 정의될 수 있다.

$$\widehat{X^2} = E\{X^2|Y\} = \frac{\int_{-\infty}^{\infty} X^2 P\{Y|X\}P\{X\}dX}{\int_{-\infty}^{\infty} P\{Y|X\}P\{X\}dX} \quad (17)$$

$A=|Y|$라고 할 때, 식 (17)로부터 다음과 같이 최소평균자승오류 추정을 얻을 수 있다[7].

$$\widehat{X^2} = \left(\frac{E\{|X|^2\}}{E\{|X|^2\}+E\{|N|^2\}}\right)^2 A^2 + \frac{E\{|X|^2\}E\{|N|^2\}}{E\{|X|^2\}+E\{|N|^2\}}. \quad (18)$$

식 (6), (7), (18)을 이용하여 최소평균자승오류에 기반을 둔 선험 SNR 추정은 다음과 같이 계산된다.

$$\xi_{MMSE} = \frac{\widehat{X^2}}{E\{|N|^2\}} = \frac{\xi}{1+\xi}\left(1+\frac{\xi}{1+\xi}\gamma\right). \quad (19)$$

선험 SNR을 추정하기 위한 첫 번째 단계는 DD 접근법을 사용하여 선험 SNR를 추정하는 것이고, 두 번째 단계는 추정된 선험 SNR을 이용하여 식 (19)에 적용, 선험 SNR을 다시 한 번 재추정하는 것이다. 즉, 위상 정보를 고려하여 추정된 선험 SNR을 최소평균자승오류 방법에 기반을 두어 재추정한 결과는 다음과 같이 계산된다.

$$\hat{\xi}_{MMSE} = \frac{\hat{\xi}}{1+\hat{\xi}}\left(1+\frac{\hat{\xi}}{1+\hat{\xi}}\tilde{\gamma}\right), \quad (20)$$

$$\tilde{\gamma} = \frac{|\tilde{Y}|^2}{E\{|\tilde{N}|^2\}}.$$

3.4.2 음성 신호 재합성

위상을 고려하여 향상된 음성 신호의 스펙트럼은 혼합 음성신호의 로그멜 스펙트럼과 잡음제거 필터(denoising filter)를 곱함으로써 계산하며, 잡음제거 필터 $H(m,k)$는 다음과 같이 정의된다.

$$H(m,k) = \frac{\hat{\xi}_{MMSE}(m,k)}{\left[1+\hat{\xi}_{MMSE}(m,k)\right]}. \quad (21)$$

잡음제거 필터를 곱하여 구한 향상된 음성 신호는 다음과 같다.

$$\hat{\tilde{X}}(m,k) = H(m,k)\tilde{Y}(m,k). \quad (22)$$

식 (22)를 이용하여 계산된 향상된 로그멜 스펙트럼을 다시 전력 스펙트럼으로 변환한 후, 역 이산 푸리에 변환(inverse discrete Fourier transform: IDFT)을 하고 오버랩-애드(overlap-add) 방법을 적용함으로써 최종적으로 전체의 향상된 음성 신호를 복원, 재합성한다.

4. 실험 결과

음성향상의 결과를 확인하고 성능을 평가하기 위하여 공개된 음성 데이터베이스에 대해 음성향상 실험을 수행하였다.

4.1 음성 데이터베이스

음성 향상 실험을 위해 사용된 음성 파일은 Interspeech 2006 음성분리대회(speech separation challenge)[8]에서 제공하는 데이터베이스에서 선택하였고, 잡음 신호는 N1(Car), N2(Babble), N3(White Gaussian)의 세 가지 종류의 잡음을 사용하였다. 혼합 음성 신호를 만들기 위하여 신호대잡음비가 -10, -5, 0, 5, 10, 15, 20 dB이 되도록 음성 신호와 잡음 신호를 혼합하였다. 실험에 사용한 음성 데이터는 여성 화자 7명과 남성 화자 7명의 각 10문장으로, 총 140 문장을 사용하였다.

음성 신호는 샘플링 주파수가 25 kHz인 음성분리대회에서 제공되는 음성 데이터를 8 kHz로 축소하여 사용하였으며, 잡음 신호는 8 kHz의 샘플링 주파수를 갖는 신호를 사용하였다. 실험에 사용된 음성 데이터는 평균 0, 분산 1을 갖도록 정규화 하였다. 정규화 된 음성 데이터를 인접한 프레임들과 16 ms가 겹치도록 하여 32 ms 크기의 프레임으로 나누어 해밍 윈도우를 적용하였다. 각 프레임에 대해 512-포인트 크기의 이산 푸리에 변환을 계산하고, 푸리에 변환 결과로부터 앞부분의 257차원 스펙트럼 벡터를 분리하여 로그멜 스펙트럼 변환에 사용하였다. 로그멜 스펙트럼 변환의 크기는 128 포인트로 정하여 사용하였다.

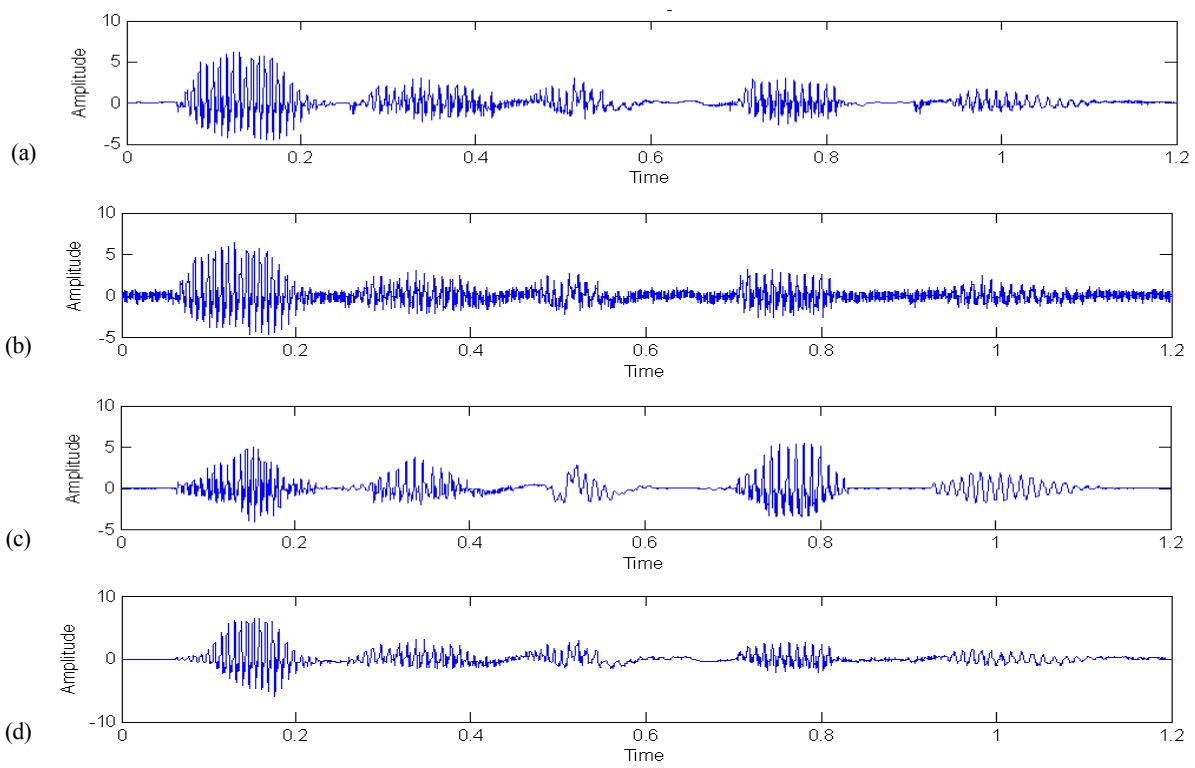

그림 3. 음성 향상 실험 파형 출력 (잡음: N3, SNR: 0 dB)
(a) 원 음성 신호, (b) 혼합 음성 신호,
(c) DD 접근법을 사용하여 향상된 음성 신호, (d) 제안된 방법을 사용하여 향상된 음성신호
Figure 3. Waveforms for
(a) original speech signal, (b) noisy signal,
(c) enhanced signal with the DD approach, (d) enhanced signal with the proposed method

4.2 음성 향상 실험 결과

혼합 음성 신호로부터 음성의 향상된 정도를 확인하기 위하여 음성신호의 파형과 스펙트로그램을 출력하였으며, 음성 향상의 결과를 수치적으로 보기 위하여 신호대잡음비를 계산하였다.

4.2.1 파형 및 스펙트로그램

그림 3은 N3 잡음을 음성 신호와 혼합한 경우의 원 음성 신호(a), 혼합된 음성 신호(b), 기존의 DD 접근법을 사용하여 향상된 음성 신호(c), 위상을 고려하여 향상시킨 신호(d)의 파형을 나타낸다. 그림 4는 그림 3의 각 파형에 대응되는 스펙트로그램을 나타낸다. 제안된 알고리듬의 성능을 측정하고 비교하기 위해 사용된 방법(Baseline)은 DD 접근법으로, 위상을 고려하지 않고 음성의 크기 정보만을 이용해 *a priori* SNR을 추정하고 음성을 향상시키는 방법이다[1].

제안된 방법을 사용하여 향상된 음성신호의 파형이 원 음성 신호와 더 가깝게 출력되었으며 잡음 요인의 신호가 눈에 띄게 제거되었음을 볼 수 있다. 스펙트로그램 출력 결과에서도 잡음 성분이 눈에 띄게 옅어지며 음성 신호의 성분이 원 음성 신호와 유사하게 출력되었고, 효과적으로 음성 향상이 수행되었음을 볼 수 있다.

4.2.2 신호대잡음비 (SNR)

표 1, 2, 3은 음성 신호의 향상된 정도를 수치적으로 확인하기 위하여 각 잡음 신호에 대해 음성 신호와 혼합된 신호로부터 기존의 크기 정보만을 이용하여 음성 향상을 수행한 경우와 제안된 알고리듬을 사용한 경우의 향상된 음성 신호의 신호대잡음비를 나타낸 것이다. 신호대잡음비는 원 음성 신호와 향상된 음성신호의 스펙트럼의 크기를 이용하여 계산하며 다음과 같다.

$$SNR = 10\log_{10}\left[\frac{|X|^2}{(|X|-|\hat{X}|)^2}\right]. \tag{23}$$

여기서 $|X|$와 $|\hat{X}|$는 각각 원 음성신호와 향상된 음성신호의 스펙트럼 크기이다. 신호대잡음비는 음성신호와 잡음에 해당하는 신호의 비를 나타낸 것으로, 신호대잡음비가 클수록 음

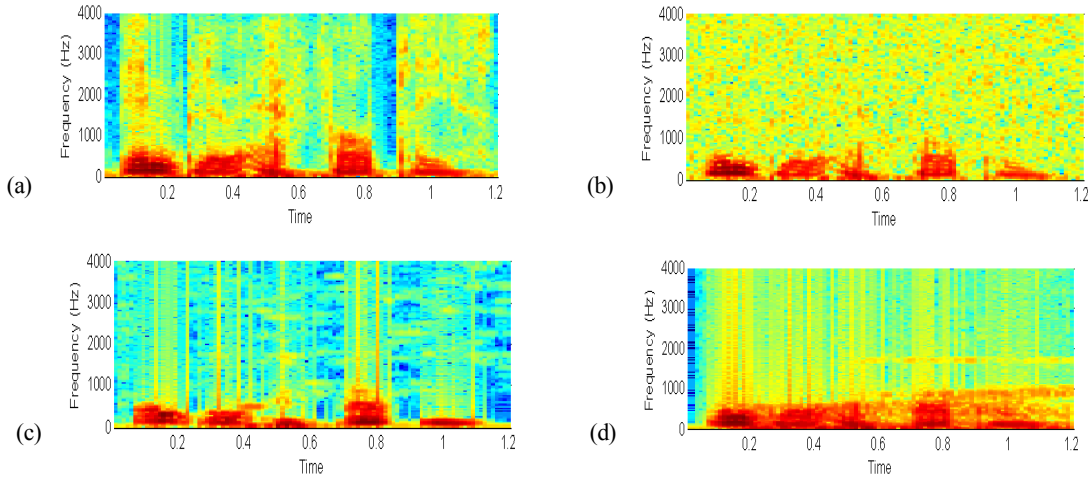

그림 4. 음성 향상 실험 스펙트로그램 출력(잡음: N3, SNR: 0 dB)
(a) 원 음성 신호, (b) 혼합 음성 신호,
(c) DD 접근법을 사용하여 향상된 음성 신호, (d) 제안된 방법을 사용하여 향상된 음성신호

Figure 4. Spectrograms for
(a) original speech signal, (b) noisy signal,
(c) enhanced signal with the DD approach, (d) enhanced signal with the proposed method

성 신호가 많고 양호한 신호라고 할 수 있다.

음성의 크기와 위상 정보를 모두 고려하여 음성 향상을 수행한 제안 알고리듬이 기존의 크기 정보만을 이용하여 음성 향상을 수행한 방법(DD 접근법)에 비해 N1의 경우 평균적으로 약 2.5 dB, N2의 경우 약 1.7 dB, 그리고 N3의 경우 약 2.6 dB의 SNR 증가로 모든 경우에서 높았으며, 전체적으로는 약 2.3 dB의 SNR 증가를 나타낸다.

표 1. 향상된 음성 신호의 평균 SNR (dB) (Car noise: N1)
Table 1. Average SNR (dB) of enhanced speech signal (N1)

입력 SNR (dB)	기존 방법	제안 방법
-10	1.5	4.4
-5	2.3	5.7
0	3.3	7.5
5	5.1	7.9
10	6.7	8.3
15	7.3	8.9
20	8.7	9.6
평균	5.0	7.5

표 2. 향상된 음성 신호의 평균 SNR (dB) (Babble noise: N2)
Table 2. Average SNR (dB) of enhanced speech signal (N2)

입력 SNR (dB)	기존 방법	제안 방법
-10	2.0	5.3
-5	4.1	6.2
0	5.9	7.9
5	6.7	8.2
10	7.0	8.5
15	7.9	9.1
20	9.0	9.7
평균	6.1	7.8

표 3. 향상된 음성 신호의 평균 SNR (dB) (White Gaussian)
Table 3. Average SNR (dB) of enhanced speech signal (N3)

입력 SNR (dB)	기존 방법	제안 방법
-10	1.1	3.9
-5	1.6	4.5
0	2.4	5.6
5	4.2	7.0
10	5.7	8.3
15	6.8	8.8
20	7.5	9.5
평균	4.2	6.8

표 4에 멜 스펙트럼의 크기가 23, 32, 64, 128일 때, 0 dB의 SNR을 가지는 혼합 음성신호의 음성향상 수행 후 계산된 평균 SNR과 PESQ를 나타내었다. 테스트 데이터는 음성 신호에 White Gaussian 잡음을 혼합하여 사용하였으며, 결과 SNR과 PESQ는 음성 향상 수행 후 계산된 SNR과 PESQ들의 평균으로 나타내었다. 로그 멜 영역에서 음성의 특징을 추출하고 음성을 향상시킨 후 다시 전력 스펙트럼 영역으로 변환하여 음성 신호를 재합성 하므로 멜 스펙트럼 크기가 23차, 32차와 같이 비교적 작은 경우 음성 신호를 재합성 하는 과정에서 신호가 뭉뚱그려져 음성 향상의 성능이 떨어지는 원인이 되며, 2.2 dB의 SNR을 보인다. 본 논문에서는 멜 스펙트럼의 크기를 성능의 개선이 크고 음성 향상 수행 후의 파형이 원 신호와 가장 가깝게 출력되는 128로 정하여 사용하였다. 기존의 DD 접근법은 5.3 dB의 SNR과 2.55의 PESQ를 보였다.

그림 5는 N3 잡음을 음성 신호와 혼합한 경우의 평균 PESQ를 나타낸 것이다. 음성의 위상은 크기 성분에 비해 사람의 귀에 상대적으로 둔감하기 때문에 성능이 크기 증가하지는 않지만 기존의 크기 성분만을 사용하여 음성을 향상한 결과에 비해 전체적으로 PESQ가 증가하며, 0 dB, 5 dB, 그리고 10 dB에서는 약 0.1의 증가를 보였다.

표 4. 멜 스펙트럼 크기에 따른 향상된 음성 신호의 평균 SNR(dB)과 PESQ(N3, 0 dB)
Table 4. Average SNR (dB) and PESQ of enhanced speech signal according to the mel-spectrum dimension (N3, 0 dB)

멜 스펙트럼 차수	제안 방법	
	SNR (dB)	PESQ
23	2.2	1.61
32	4.1	2.50
64	7.7	2.71
128	7.9	2.75

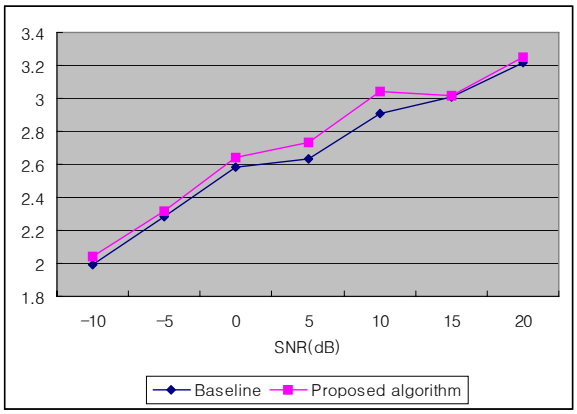

그림 5. 향상된 음성 신호의 평균 PESQ(잡음: N3)
Figure 5. Average PESQ of enhanced speech signal (N3)

5. 결론

본 논문에서는 음성의 크기와 위상 성분을 모두 고려하여 로그멜 영역에서 선험 SNR을 추정하고 음성을 향상시키는 단일 채널 음성 향상 시스템을 제안하였다. 제안한 방법에서는 위상 구간을 평균적으로 0으로 만들지 않기 위해 혼합 음성신호의 전력 스펙트럼 벡터를 로그멜 스펙트럼 벡터로 변환한 후 위상 기반의 선험 SNR을 추정한다. 추정된 선험 SNR을 최소평균자승오류 방법에 기반을 두어 다시 한 번 재추정한 후 이를 이용하여 원하는 음성 신호를 추정하고 음성 신호를 향상시킴으로써 하나의 시간 프레임 지연에서 후험 SNR의 형태를 따라가는 선험 SNR의 문제를 해결하고 음성 신호의 크기 정보 뿐 아니라 위상 정보를 함께 고려하여 음성 신호 향상의 성능을 높였다.

공개 음성 데이터베이스를 사용하여 음성 신호 향상 실험을 수행한 결과, 제안된 방법을 사용하여 향상된 음성 신호의 파형과 스펙트로그램 출력이 원 신호에 가깝게 출력되었고, 잡음 요인의 신호가 눈에 띄게 제거되었다. 신호대잡음비를 계산하여 정량적으로 성능을 측정한 결과에서도 기존의 크기 정보만을 이용한 경우에 비하여 전체적으로 약 2.3 dB의 SNR 증가를 보였다.

참고문헌

Ephraim, Y., Malah, D. (1984). “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator”, *IEEE Trans. Acoustics, Speech and Signal Processing,* Vol. 32, pp. 1109-1121.

Loizou, P. C. (2007). *Speech Enhancement,* CRC Pr I Llc, pp. 97-289.

Alam, M. J., O’Shaughnessy, D., Selouani, S.-A. (2008). “Speech enhancement based on novel two-step a priori SNR estimators”, *Proc. Interspeech*, pp. 565-568, Sep.

Faubel, F., McDonough, J., Klakow, D. (2008). “A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain”, *Proc. Interspeech*, pp. 553-556, Sep.

Paliwal, K. K., (2003). “Usefulness of phase in speech processing”, *Proc. IPSJ Spoken Language Processing Workshop,* Gifu, Japan, pp. 1-6.

Lee, Y.-K., Kwak, C., Lee, I. S., Kwon, O.-W. (2010). “Single-channel speech separation using zero-phase models”, *IEEE International Symposium on Consumer Electronics*, pp. 1-5.

Accardi, A. J., Cox, R. V. (1999). “A modular approach to speech

enhancement with an application to speech coding", *in Proc. IEEE Int. Conf. Acoustic, Speech, and Signal Processing,* Vol. 1, pp. 201-204.

Cooke, M., Lee, T.-W. (2010) "Speech Separation and Recognition Competition", http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.

• **이윤경 (Lee, Yun-Kyung)**
충북대학교 제어로봇공학과
충북 청주시 흥덕구 성봉로 410 (개신동)
Tel: 043-261-3374
Email: yklee@cbnu.ac.kr
관심분야: 음원분리, 음성인식, 반향제거, 음성 및 오디오 처리
현재 제어로봇공학과 대학원 박사과정 재학 중

• **권오욱 (Kwon, Oh-Wook)** 교신저자
충북대학교 전자정보대학
충북 청주시 흥덕구 성봉로 410 (개신동)
Tel: 043-261-3374
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 음성 및 오디오 처리, 패턴 인식
2003~현재 전자공학과 부교수