한국음향학회지 제27권 제7호 pp. 379-385 (2008)

원단 잡음 환경에서 Soft Decision에 기반한 새로운 음성 강화 기법

Speech Reinforcement Based on Soft Decision Under Far-End Noise Environments

최 재 훈*, 장 준 혁*
(Jae-Hun Choi*, Joon-Hyuk Chang*)

*인하대학교 전자공학부
(접수일자: 2008년 5월 30일; 채택일자: 2008년 7월 3일)

본 논문에서는 근단 (Near-End)및 원단 (Far-End) 잡음 환경에서 효과적인 음성 강화 기법을 제시한다. 일반적으로 배경 잡음이 존재하는 근단 환경에서 수신하는 원단 화자 음성의 명료도가 매우 감소하므로, 이를 극복하기 위한 원단 화자 음성 강화 기법이 필요하다. 구체적으로, 추정된 근단 화자의 배경 잡음 전력을 기반으로 원단 화자의 음성 전력을 강화시키는데, 특별히 근단 환경에서도 잡음이 존재하는 일반적인 경우를 고려하여, 잡음에 오염된 원단 음성 신호중 잡음을 제외한 실제 음성 신호만 강화하는 개선된 알고리즘을 제안한다. 제안된 음성 강화 기법의 성능은 다양한 잡음 환경 하에서 ITU-T P.800의 주관적 음질 측정 방법인 CCR (Comparison Category Rating) 테스트에 의해 평가되었으며, 기존의 음성 강화 기법과 비교해서 우수한 성능을 보여주었다.

핵심용어: Soft Decision, 마스킹 효과, SNR 복구, 근단 배경 잡음, 음성 강화
투고분야: 음성처리 분야 (2,3)

In this paper, we propose an effective speech reinforcement technique under the near-end and the far-end noise environments. In general, since the intelligibility of the far-end speech for the near-end listener is significantly reduced under near-end noise environments, we require a far-end speech reinforcement approach to avoid this phenomena. Specifically, based on the estimated background noise spectrum of the near-end, we reinforce the far-end speech spectrum by incorporating the more general cases under the near-end with background noise. Also, we propose the novel approach to reinforce the actual speech signal except for the noise signal in the far-end noisy speech signal. The performance of the proposed algorithm is evaluated by the CCR (Comparison Category Rating) test of the method for subjective determination of transmission quality in ITU-T P.800 under various noise environments and shows better performances compared with the conventional method.

Keywords: Soft Decision, Masking effect, SNR Recovery, Nbackground noise, Speech reinforcement
ASK subject classification: Speech Signal Processing (2,3)

I. 서 론

다양한 배경 잡음이 존재하는 환경에서 휴대폰을 이용해 상대방과 음성 통화를 하게 되면 배경 잡음으로 인해 상대방 음성이 잘 들리지 않게 된다. 예를 들어 지하철과 같은 소음이 큰 곳에서 휴대폰으로 통화를 하거나, 자동차가 다니는 길거리에서 통화 시에 소음이 없는 조용한 곳에서 통화하는 것보다 상대방의 음성이 매우 작게 들리게 된다. 이럴 경우 우리는 의식적으로 상대방의 음성을 잘 들을 수 있도록 휴대폰의 볼륨을 최대로 키우는 노력을 하게 되는데, 이러한 현상은 심리 음향학 용어로 마스킹 효과 (Masking effect)라고 한다 [3].

마스킹 효과 (Masking effect)란 특정 큰 신호 (Masker)가 존재할 때 함께 존재하는 다른 작은 신호 (Maskee)는 전혀 들리지 않게 되는 것을 말하며, 마스킹 효과 중에서

책임저자: 장 준 혁 (changjh@inha.ac.kr)
인천시 남구 용현동 253 인하대학교 전자공학부
(전화: 032-860-7423; 팩스: 032-868-3654)

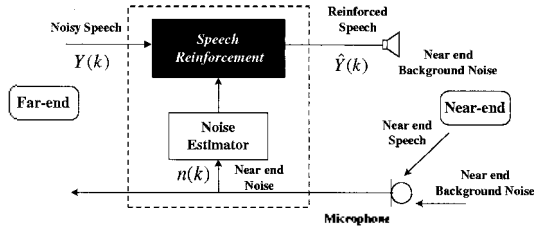

그림 1. 추정된 근단 화자의 배경 잡음 전력에 기반한 제안된 음성 강화 시스템 개략도

Fig. 1. Block diagram of the proposed speech reinforcement system based on the estimated near-end listener's background noise spectrum.

도 부분 마스킹 효과 (Partial masking effect)에 의하면 비슷한 크기의 두 신호가 존재할 때 두 신호 모두 원래 신호 크기보다 작게 들리게 된다. 서로 부분적으로 마스킹 하게 되는데 이에 따라 잡음이 존재할 경우 함께 존재하는 음성의 크기는 작게 들리게 된다.

기존의 음성 향상 기법은 원단에서 음성을 전송하기 전에 노이즈 부분의 제거에 초점을 맞추어 왔다. 그러나 배경 잡음이 존재하는 근단 환경에서는 배경 잡음 자체가 근단 화자의 귀에 직접적으로 전달되기 때문에 노이즈가 제거된 원단 음성 신호가 근단에 도달하여도 근단의 배경 잡음으로 인해 원단 음성 신호의 명료성은 저하되게 된다. 그러므로 근단의 배경 잡음이 존재하는 상황에서 근단 화자가 듣게 되는 원단 음성 신호의 명료도를 향상시키기 위해 원단 음성 신호를 강화하는 기법이 필요하다.

이러한 배경 잡음이 근단 화자에 미치는 영향에 대한 최신의 연구로 음성의 명료성 (Speech Intelligibility)에 관한 Sauert와 Vary의 주파수별 SNR 복구 기법으로 주파수별 신호 대 잡음 비 (Signal-to-Noise Ratio, SNR)가 일정하게 유지되도록 전송되어 온 음성을 증폭시키는 방법이 있다 [4]. 이 기법은 원단으로부터 전송되어 온 음성 신호의 모든 주파수 성분에 똑같은 이득을 곱해줌으로써 음질과 명료도를 향상시키는 것이다. 그러나 이 기법은 원단 음성 신호가 잡음이 없는 깨끗한 음성 신호라 가정하기 때문에, 잡음이 존재하는 일반적인 통신 환경을 고려한다면, 실제 상황에 적용하기에는 상당한 문제점을 가지고 있다. 구체적으로 원단의 오염된 음성 신호가 근단에 전송되었을 경우 음성과 함께 잡음까지도 증폭시켜주게 되어, 근단 화자가 듣게 되는 음성 신호의 음질과 명료도가 더 나빠지는 결과를 초래하게 된다.

본 논문에서는, 다양한 배경 잡음이 존재하는 근단 환경에서 추정된 근단 화자의 배경 잡음 전력을 기반으로 오염된 원단 화자의 음성 전력 중에서 잡음 구간을 제외한 실제 음성이 존재하는 구간에서 잡음이 섞이지 않은 깨끗한 음성 신호 전력만 추정하여, 증폭을 시킴으로써 근단 화자에 전송된 오염된 원단 음성 신호의 명료도를 향상시키는 새로운 음성 강화 기법을 제시한다. 그림 1은 제안된 음성 강화 기법의 전체적인 모습을 설명하는 개략도를 나타낸다. 기존의 SNR 복구 기법에서 원단 음성 신호의 short-term power spectral density (PSD)를 추정함에 있어 반복된 실험을 통해 얻어진 고정된 시간 상수를 사용하였다. 반면에 본 논문에서 제안하는 음성 강화 기법에서는 Soft Decision을 기반으로 통계적인 신뢰성을 갖는 음성부재확률 (SAP)을 결합함으로써 음성의 존재구간과 잡음구간을 구별하고, 여기서 구해진 음성존재 확률을 오염된 원단 음성 신호에 적용하여, 잡음이 섞이지 않은 깨끗한 음성만을 추정하였다. 추정된 깨끗한 음성 신호를 근단에서 추정된 배경 잡음 전력과의 연산을 통해 이득을 구하고, 오염된 원단 음성 신호중에서 깨끗한 음성 신호 전력만을 증폭함으로써 근단에 존재하는 배경 잡음으로 인해 음성의 명료도가 떨어진 원단 음성 신호의 명료도를 향상시켰다.

제시된 알고리즘의 성능은 다양한 배경 잡음이 존재하는 환경 하에서 ITU-T P.800의 표준 음질 측정 방법인 CCR (Comparison Category Rating) 테스트로 실험을 진행하였으며, 기존의 SNR 복구 기법보다 향상된 결과를 나타내었다 [10].

II. 주파수별 SNR 복구 기법의 개요

Sauert와 Vary는 배경 잡음이 존재하는 근단 환경에서 원단으로부터 전송된 음성 신호의 음질과 명료성을 향상시키기 위해 주파수별 SNR 복구 알고리즘을 제시하였다 [4]. 시간축 상에서 원래의 깨끗한 음성 신호 $s(t)$와 근단에 존재하는 배경 잡음 신호 $n(t)$을 DFT (Discrete Fourier Transform)를 통해 주파수 축으로 변환하면 각각 $S(t,k)$와 $N(t,k)$로 나타낼 수 있다. 여기서, $S(t,k)$, $N(t,k)$는 t번째 프레임에서의 k번째 주파수 성분을 나타낸다. 음성 신호를 이득 $G(t,k)$에 의해 증폭된 신호로 표현하면 다음과 같이 나타낼 수 있다.

$$\hat{S}(t,k) = G(t,k) \cdot S(t,k) \tag{1}$$

원단의 깨끗한 음성 신호의 short-term power spectral

density (PSD)를 $\Phi_{SS}(t,k)$라 하고, 이득 $G(t,k)$에 의해 증폭된 근단 음성 신호의 short-term (PSD)를 $\Phi_{s_1s_1}(t,k)$로 나타낼 수 있다. 따라서 증폭된 근단 음성 신호의 short-term (PSD) $\Phi_{s_1s_1}(t,k)$와 근단의 배경 잡음 전력 short-term (PSD) $\Phi_{NN}(t,k)$의 비는 SNR $\xi = 15$ dB와 같거나 더 커야하며, 다음과 같이 표현된다 [4].

$$\frac{\Phi_{s_1s_1}(t,k)}{\Phi_{NN}(t,k)} \geq \xi \tag{2}$$

이득 $G(t,k)$은 결정되어져 있기 때문에 식 (2)은 다음과 같다.

$$\frac{\Phi_{s_1s_1}(t,k)}{\Phi_{NN}(t,k)} = \frac{G^2(t,k) \cdot \Phi_{SS}(t,k)}{\Phi_{YY}(t,k)} \geq \xi \tag{3}$$

또한, 식(3)을 이득 $G(t,k)$에 의한 식으로 나타내면 다음과 같은 식으로 나타낼 수 있다.

$$G(t,k) \geq \sqrt{\xi \cdot \frac{\Phi_{NN}(t,k)}{\Phi_{SS}(t,k)}} \tag{4}$$

원단의 음성 신호는 근단의 배경 잡음 환경에서 감쇄되지 않아야 하기 때문에 다음과 같이 이득에 대한 제약 조건을 도입할 수 있다.

$$G(t,k) \geq 1 \tag{5}$$

따라서 이득 $G(t,k)$을 (4)식과 (5)식의 제약 조건과 결합하면 아래와 같이 표현된다.

$$G(t,k) = \max\left\{\sqrt{\xi \cdot \frac{\Phi_{NN}(t,k)}{\Phi_{SS}(t,k)}}, 1\right\} \tag{6}$$

근단에 도달한 원단 음성 신호가 과도하게 증폭되지 않도록 이득 $G(t,k)$을 최대 이득 $G_{\max} \cong 30dB$로 제한한다. 따라서 (6)식과 결합하면 다음과 같다.

$$G(t,k) = \min\left\{\max\left\{\sqrt{\xi \cdot \frac{\Phi_{NN}(t,k)}{\Phi_{SS}(t,k)}}, 1\right\}, G_{\max}\right\} \tag{7}$$

원단에서 전송된 깨끗한 음성 신호의 short-term PSD $\Phi_{SS}(t,k)$와 근단의 배경 잡음전력의 short-term PSD $\Phi_{NN}(t,k)$은 신호의 스펙트럼 전력 밀도로 다음과 같이 계산되어진다.

$$\begin{aligned}\Phi_{SS}(t,k) &= \alpha_s \cdot \Phi_{SS}(t-1,k) + (1-\alpha_s) \cdot |S(t,k)|^2 \\ \Phi_{NN}(t,k) &= \alpha_N \cdot \Phi_{SS}(t-1,k) + (1-\alpha_N) \cdot |N(t,k)|^2\end{aligned} \tag{8}$$

short-term PSD $\Phi_{SS}(t,k)$와 short-term PSD $\Phi_{NN}(t,k)$을 계산하는데 사용된 α_s, α_N는 실험적으로 최적화된 고정 시간 상수로써 $\alpha_s = 0.996$, $\alpha_N = 0.96$이다 [4].

III. Soft Decision에 기반한 원단 음성 강화 기법

제안된 음성 강화 시스템의 전체 블록 도를 그림 2에 나타내었는데, 잡음이 섞인 오염된 음성 신호에서 잡음 신호와 실제 음성 신호를 구분하기 위해 Soft Decision에 기반한 음성 존재와 음성 부재의 확률을 적용하여, 오염된 원단 음성 신호에서 잡음 신호 전력과 실제 음성 신호 전력을 정확하게 추정하는 것이다. 이를 바탕으로 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정하여, 근단의 배경 잡음 전력과의 연산을 통해 얻어진 이득만큼 추정된 깨끗한 음성 신호만을 증폭하여 근단 화자가 듣게 되는 오염된 음성 신호의 명료성을 강화 시키는 것이다. 따라서 원단에서 입력되는 오염된 음성 신호에서 깨끗한 음성 신호 전력의 정확한 추정이 가장 중요하다.

원단에서의 깨끗한 음성 신호 $s(t)$과 잡음 신호를 $d(t)$라 한다면 오염된 음성 신호를 $y(t)$으로 나타낼 수 있으며, 각각의 성분을 DFT (Discrete Fourier Transform)을 통해서 주파수 축으로 나타내면 다음과 같이 나타낼 수 있다.

$$Y(t,k) = S(t,k) + D(t,k) \tag{9}$$

여기서 $Y(t,k)$, $S(t,k)$, $D(t,k)$는 $y(t)$, $s(t)$, $d(t)$의

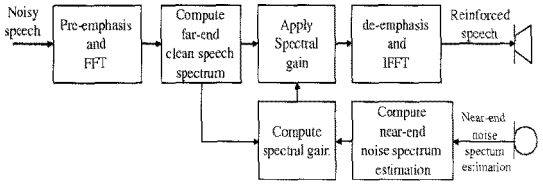

그림 2. 제안된 음성 강화 기법의 전체 블록도
Fig. 2. Overall block diagram of the proposed speech reinforcement approach.

t번째 프레임에 대한 k번째 주파수 성분을 표시한다. 추정된 근단의 배경 잡음 전력과 오염된 원단 음성 신호에서 깨끗한 음성 신호 전력의 추정치로 구해진 이득 $G(t,k)$을 곱해서 강화된 음성 신호 $\hat{y}(t)$을 다음과 같이 나타낼 수 있다.

$$\hat{Y}(t,k) = G(t,k) \cdot Y(t,k) \tag{10}$$

추정된 근단 화자의 배경 잡음 전력과 오염된 원단 음성 신호에서 잡음 신호를 제외한, 깨끗한 음성 신호 전력의 추정을 통하여 이득 $G(t,k)$에 대한 (7)식을 다음과 같이 다시 나타낼 수 있다.

$$G(t,k) = \min\left\{\max\left\{\sqrt{\xi \cdot \frac{\hat{\Phi}_{NN}(t,k)}{\hat{\Phi}_{SS}(t,k)}}, 1\right\}, G_{\max}\right\} \tag{11}$$

식 (11)에서 나타낸 $\hat{\Phi}_{SS}(t,k)$는 오염된 원단 음성 신호에서 잡음이 섞이지 않은 깨끗한 음성 신호 전력 $\Phi_{SS}(t,k)$의 추정 값을 나타낸다. 기존의 Sauert와 Vary의 주파수별 SNR 복구 기법에서는 원단의 깨끗한 음성 신호 전력 $\Phi_{SS}(t,k)$을 구함에 있어, 반복된 실험에 의해 구해진 고정된 시간 상수 α_s을 적용하였다. 또한 원단에 입력된 음성 신호가 오염된 음성 신호일 경우, 잡음 신호와 음성 신호가 섞여 있게 되므로, 잡음 신호와 음성 신호 모두를 증폭시키는 문제점을 가지고 있다. 그러나 제안된 음성 강화 기법에서는 오염된 음성 신호로부터 잡음 신호와 실제 음성 신호를 정확하게 추정하고, 잡음 신호를 제외한 추정된 깨끗한 음성 신호 전력만을 증폭시키게 된다. 따라서 식(11)의 $\hat{\Phi}_{SS}(t,k)$는 오염된 원단 음성 신호 중에서 실제 잡음이 섞이지 않는 깨끗한 음성 신호 전력만을 추정한 값을 나타낸다. 오염된 원단 음성 신호의 깨끗한 음성 신호 전력의 추정 값 $\hat{\Phi}_{SS}(t,k)$는 Soft Decision에 기반하여 다음과 같이 구할 수 있다.

원단 신호의 음성 부재와 음성 존재에 대한 가설을 각각 H_0, H_1이라고 하면, 주파수 채널에 따라 다음과 같이 가정할 수 있다.

$$\begin{aligned} H_0 &: \textit{speech absence} : Y(t) = D(t) \\ H_1 &: \textit{speech presence} : Y(t) = S(t) + D(t) \end{aligned} \tag{12}$$

음성 신호와 잡음 신호의 스펙트럼이 zero-mean 복소 가우시안 분포를 보인다고 가정하면, (12)에서 제시한 가설 H_0, H_1에 따라 다음과 같은 확률 밀도 함수로 표현할 수 있다.

$$\begin{aligned} p(Y(t,k)|H_0) &= \frac{1}{\pi\lambda_d(t,k)}\exp\left\{-\frac{\Phi_{YY}(t,k)}{\lambda_d(t,k)}\right\} \\ p(Y(t,k)|H_1) &= \frac{1}{\pi[\lambda_d(t,k)+\lambda_s(t,k)]} \\ &\quad \cdot \exp\left\{-\frac{\Phi_{YY}(t,k)}{\lambda_d(t,k)+\lambda_s(t,k)}\right\} \end{aligned} \tag{13}$$

위의 식에서 $\lambda_s(t,k)$과 $\lambda_d(t,k)$는 각각 t번째 프레임에 대한 k번째 주파수 성분에서의 음성과 잡음의 분산을 나타낸다. 따라서 음성의 존재와 부재에 관한 가설로부터 주파수 채널별 음성부재확률인 LSAP는 다음과 같다.

$$\begin{aligned} p(H_0|Y(t,k)) &= \frac{p(Y(t,k)|H_0)p(H_0)}{p(Y(t,k))} \\ &= \frac{p(Y(t,k)|H_0)p(H_0)}{p(Y(t,k)|H_0)p(H_0)+p(Y(t,k)|H_1)p(H_1)} \\ &= \frac{1}{1+\frac{p(H_1)}{p(H_0)}A(Y(t,k))} \end{aligned} \tag{14}$$

식에서 $p(H_0)$는 음성 부재에 대한 사전 확률 (*a priori* probability)이고, $A(Y(t,k))$는 k번째 주파수 대역의 우도비 (likelihood ratio)로써 다음과 같이 표현된다.

$$\begin{aligned} A(Y(t,k)) &= \frac{p(Y(t,k)|H_1)}{p(Y(t,k)|H_0)} \\ &= \frac{1}{1+\xi(t,k)}\exp\left[\frac{\gamma(t,k)\xi(t,k)}{1+\xi(t,k)}\right] \end{aligned} \tag{15}$$

(15)식에서 $\gamma(t,k)$, $\xi(t,k)$는 각각 *a posteriori* SNR과 predicted SNR로써 다음과 같이 나타내어진다.

$$\begin{aligned} \gamma(t,k) &= \frac{\Phi_{YY}(t,k)}{\Phi_{DD}(t,k)} \\ \xi(t,k) &= \frac{\Phi_{SS}(t,k)}{\Phi_{DD}(t,k)} \end{aligned} \tag{16}$$

새롭게 제안된 음성 강화 기법의 알고리즘의 경우 오염된 원단 음성 신호 중에서 잡음이 섞이지 않은 깨끗한 음성 신호만이 증폭의 대상이기 때문에, 잡음 전력 $\Phi_{DD}(t,k)$과 음성 전력 $\Phi_{SS}(t,k)$의 추정이 알고리즘 성능

에 중요한 역할을 하게 된다. 일반적으로, 음성검출기(VAD, voice activity detector)를 사용하여 음성 부재구간에서 잡음 전력을 갱신한다 [6]. 그러나 실제 잡음 환경이 비정상 (non-stationary)인 경우, 잡음 전력은 음성의 존재 구간에서도 갱신되어야 한다. 따라서 신뢰성 있는 음성 전력 $\Phi_{SS}(t,k)$와 잡음 전력 $\Phi_{DD}(t,k)$을 추정하기 위해, 음성 전력과 잡음 전력 각각에 long-term smoothed 전력 스펙트럼을 사용하여 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\hat{\Phi}_{SS}(t+1,k) &= \zeta_s \hat{\Phi}_{SS}(t,k) + (1-\zeta_s)\Phi_{SS}(t,k) \\ \hat{\Phi}_{DD}(t+1,k) &= \zeta_d \hat{\Phi}_{DD}(t,k) + (1-\zeta_d)\Phi_{DD}(t,k)\end{aligned} \tag{17}$$

식(17)에서 $\hat{\Phi}_{DD}(t,k)$와 $\hat{\Phi}_{SS}(t,k)$는 각각 $\Phi_{DD}(t,k)$과 $\Phi_{SS}(t,k)$의 추정 값이고, ζ_d과 ζ_s는 정상 (stationary) 상태를 가정한 스무딩 파라미터로써 $0 < \zeta_d, \zeta_s < 1$ 값을 가진다. 음성 신호 $S(t)$와 잡음 신호 $D(t)$의 통계적 가정과 (17)식을 이용하여 현재 프레임에 대한 음성 신호와 잡음 신호의 전력 추정 값을 계산하면 다음 식으로 나타낼 수 있다.

$$\begin{aligned}\Phi_{SS}(t,k) = {} & E\left[|S(t,k)|^2 | Y(t,k), H_0\right] p(H_0 | Y(t,k)) \\ & + E\left[|S(t,k)|^2 | Y(t,k), H_1\right] p(H_1 | Y(t,k))\end{aligned} \tag{18}$$

$$\begin{aligned}\Phi_{DD}(t,k) = {} & E\left[|D(t,k)|^2 | Y(t,k), H_0\right] p(H_0 | Y(t,k)) \\ & + E\left[|D(t,k)|^2 | Y(t,k), H_1\right] p(H_1 | Y(t,k))\end{aligned} \tag{19}$$

여기서,

$$\begin{aligned}& E\left[|D(t,k)|^2 | Y(t,k), H_0\right] = \Phi_{YY}(t,k) \\ & E\left[|D(t,k)|^2 | Y(t,k), H_1\right] \\ & = \left(\frac{\hat{\xi}(t,k)}{1+\hat{\xi}(t,k)}\right)\hat{\Phi}_{DD}(t,k) + \left(\frac{1}{1+\hat{\xi}(t,k)}\right)^2 \Phi_{YY}(t,k)\end{aligned} \tag{20}$$

$$\begin{aligned}& E\left[|S(t,k)|^2 | Y(t,k), H_0\right] = 0 \\ & E\left[|S(t,k)|^2 | Y(t,k), H_1\right] \\ & = \left(\frac{1}{1+\hat{\xi}(t,k)}\right)\hat{\Phi}_{SS}(t,k) + \left(\frac{\hat{\xi}(t,k)}{1+\hat{\xi}(t,k)}\right)^2 \Phi_{YY}(t,k)\end{aligned} \tag{21}$$

위에서 사용된 $\hat{\xi}(t,k)$는 predicted SNR로 정의되며, 추정된 잡음 전력과 음성 전력의 비로 다음과 같이 나타내어진다.

$$\hat{\xi}(t,k) = \frac{\Phi_{SS}(t,k)}{\Phi_{DD}(t,k)} \tag{22}$$

predicted SNR은 잡음 전력 갱신을 위해 사용된다.

IV. 실험 방법

본 논문에서는 제안된 음성 강화 알고리즘의 성능을 평가하기 위해서 ITU-T P.800의 표준 음질 측정 방법인 CCR (Comparison Category Rating) 테스트로 실험을 실시하였다 [10]. 제안된 음성 강화 알고리즘은 추정된 근단 화자의 배경 잡음 전력을 기반으로 오염된 원단 음성 신호중 잡음을 제외한 실제 음성 구간을 증폭시켜서 배경 잡음이 존재하는 근단 화자가 듣게 되는 원단 음성 신호의 명료성 향상을 목적으로 한다. 따라서 원단 음성 신호가 오염된 신호임을 가정하기 위해, 남성과 여성 화자가 각각 10개의 문장을 발음한 음성을 8 khz로 샘플링 한 후에 NOISEX-92 데이터베이스의 white gaussian noise (WGN)을 SNR 10dB로 부가하였다. 실제 배경 잡음이 존재하는 환경에서 실험을 진행하는 대신에 오염된 원단 음성 신호와 강화된 근단 음성 신호에 근단의 배경 잡음으로써 NOISEX-92 데이터베이스의 white gaussian noise (WGN), babble와 vehicle을 잡음레벨이 SNR 5 dB, 10 dB, 15 dB, 20 dB로 섞이도록 하였고, 헤드폰을 통해 들을 수 있게 실험을 진행하였다.

총 14명의 청자들에 의해 주관적 테스트 방법인 CCR 테스트가 진행되었으며, 14명의 청자들은 3가지 종류의 음성 파일을 듣게 된다. 첫 번째 파일은 reference 파일로써 잡음이 없는 깨끗한 음성 파일을 참고용으로 듣게 된다. 그리고 14명의 청자는 2개의 파일을 각각 듣게 되는데, 하나의 파일은 제안된 알고리즘이 적용된 파일이고, 다른 하나의 파일은 제안된 알고리즘과 비교 대상이 되는 파일이다. 이때 참고용 reference파일을 제외한 비교 대상 2개의 파일은 순서 없이 재생된다. reference 파일을 기준으로 2개의 파일 중 어떤 음성 신호가 얼마나 음질이 좋은지, 배경 잡음에 대해 음성 신호가 얼마나 명료하게 들리는지, 얼마나 크게 잘 들리는지, 왜곡은 없는지를 평가 기준으로 -3점부터 +3점까지 주관적인 점수를 매기게 된다. 표에 기록된 총 점수는 평균값으로 나타내었는데, 결과에 나타난 점수가 0보다 큰 양수 값을 가질수록 제안된 알고리즘이 성능이 더 좋다는 것을 의미하고, 반대로 0보다 작은 음수의 평균값을 가지게 된다면 강화된

음성 신호의 성능이 나쁘다는 것을 의미하게 된다.

첫 번째 실험의 목적은 근단 배경 잡음이 존재하는 환경에서 근단 화자가 오염된 원단 음성을 들었을 경우, 어떠한 음성 처리 기법도 적용되지 않고 전송된 음성 신호 대비 제안된 음성 강화 기법에 의해 강화 원단 음성 신호가 얼마나 더 명료하게 들리는가를 평하기 위한 것이다. 표 1에 나타난 결과에 따르면 모든 SNR에서 제시된 음성 강화 알고리즘을 적용하여 강화된 음성 신호가 어떠한 음성 처리 기법도 거치지 않고 근단에 전송된 오염된 원단 음성 신호 대비 평균값이 0보다 큰 양수 값을 가짐을 볼 수 있는데, 이를 통해 제안된 음성 강화 알고리즘의 성능이 우수함을 알 수 있다.

두 번째 실험은 기존의 Sauert와 Vary가 제안한 SNR 복구 알고리즘이 적용된 오염된 원단 음성 신호 대비 제안된 음성 강화 알고리즘이 적용된 오염된 원단 음성 신호와의 비교를 통해 근단 화자가 듣게 되는 음성의 명료성을 평가하기 위한 것이다. 표 2에 나타난 결과를 보면 babble, vehicle, white noise에서 모두 0보다 큰 양수의 평균값을 나타냄을 볼 수 있다. 이것은 기존의 SNR 복구 알고리즘 방법보다 제안된 음성 강화 알고리즘이 더 우수한 성능을 나타냄을 보여준다. 그러나 표 2에 나타난 점수가 일관된 결과를 보여주지 못하고 있는데, 이것은 잡음의 주파수 성격에 따라 음성 강화의 효과가 다르기 때문이라고 추정된다.

제안된 음성 강화 알고리즘이 기존의 SNR 복구 기법보다 더 우수한 성능을 보인다는 것은 그림 3을 통해서도 볼 수 있다. 그림 3을 보면 (a)은 오염된 원단 음성 신호를 나타내고, (b)는 SNR 복구 기법에 의해 음성과 잡음 구간이 모두 증폭된 음성 신호 파형을 나타내고 (c)는 제안된 음성 강화 기법에 의해 잡음이 섞이지 않은 깨끗한 음성 신호 전력만 증폭시킨 것으로, 기존의 방법과 비교해 음질 향상을 확인할 수 있었다.

표 1. 다양한 배경 잡음 환경에서 음성 처리 알고리즘이 적용되지 않은 경우 대비 제안된 음성 강화 알고리즘에 대한 CCR 테스트 결과 (95% 신뢰 구간)

Table 1. The CCR results for the proposed reinforced algorithm with respect to the unprocessed under various background noise (With 95% confidence interval).

noise	SNR (dB)	Scores
white	5	1.49 ± 0.14
	10	1.46 ± 0.16
	15	1.36 ± 0.19
	20	1.33 ± 0.20
babble	5	1.49 ± 0.14
	10	1.50 ± 0.15
	15	1.48 ± 0.19
	20	1.55 ± 0.18
vehicle	5	1.61 ± 0.14
	10	1.87 ± 0.15
	15	1.70 ± 0.19
	20	1.68 ± 0.21

표 2. 다양한 배경 잡음 환경에서 SNR 복구 알고리즘 대비 제안된 음성 강화 알고리즘에 대한 CCR 테스트 결과 (95% 신뢰구간)

Table 2. The CCR results for the proposed reinforced algorithm with respect to the SNR Recovery algorithm under various background noise (With 95% confidence interval).

noise	SNR (dB)	Scores
white	5	0.81 ± 0.17
	10	0.87 ± 0.18
	15	1.12 ± 0.20
	20	1.09 ± 0.23
babble	5	0.46 ± 0.17
	10	0.41 ± 0.18
	15	0.39 ± 0.20
	20	0.31 ± 0.20
vehicle	5	0.00 ± 0.15
	10	0.19 ± 0.15
	15	0.10 ± 0.16
	20	0.17 ± 0.19

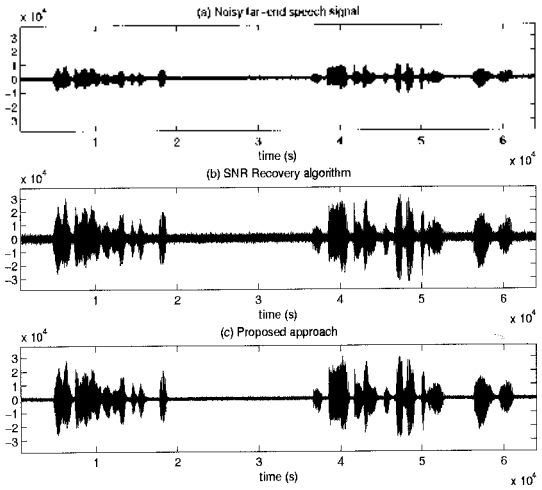

그림 3. 음성신호의 비교 (a) 오염된 원단음성신호 (b) 기존의 SNR 복구 기법에 의해 음성과 잡음구간 모두 증폭된 음성 신호 (c) 제안된 음성 강화 기법에 의해 음성 구간만 증폭된 음성 신호 [4]

Fig. 3. Comparisons of the speech signal. (a) Noisy far end speech signal (b) the speech signal based on the SNR Recovery algorithm (c) the speech signal based on the reinforcement approach [4]

V. 결 론

본 논문에서는 근단 배경 잡음이 존재하는 환경 하에서, 추정된 근단 화자의 배경 잡음 전력을 기반으로 오염된 원단 음성 신호 전력을 강화하는 새로운 알고리즘을 제안하였다. 잡음에 오염된 원단 음성 신호 전력에서 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정함에 있어, Soft Decision에 기반한 통계적인 신뢰성을 갖는 채널별 음성부재확률 (SAP)을 결합하여 정확한 추정치를 구할 수 있었고, 이를 바탕으로 오염된 원단 음성 신호로부터 실제 음성 신호만 추정하여, 추정된 원단의 깨끗한 음성 신호만을 강화함으로써 다양한 배경 잡음이 존재하는 곳에서 근단 화자가 듣게 되는 오염된 원단 음성 신호의 명료성을 향상시킬 수 있었다. 실험 결과는 제안된 음성 강화 알고리즘이 기존에 제시된 SNR 부구 알고리즘보다 더 우수한 결과를 보여준다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업 [2008-F-045-01]과 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (ITTA-2008-C1090-0804-0007).

참 고 문 헌

1. N. S. Kim, J. -H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Processing Letters, 7(5), May 2000, pp. 108-110.
2. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., ASSP-32(6), 1109-1121, Dec. 1984.
3. B. C. J. Moore, *An Introduction to the Psychology of Hearing*, (Academic Press, 2003).
4. B. Sauert and P. Vary, "Near end listening enhancement :Speech intelligibility improvement in noisy environmnets," in Proc. IEEE Int. Conf. Acoustics., Speech, Signal Processing., 1(I-493-I-496), 2006.
5. O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process., 2(2), 345-349, Apr. 1994.
6. J. Sohn, N. S. Kim, W. Sung, "A statistical model-based voice activity detection," IEEE Signal Processing Letters, 6(1), 1-3, Jan. 1999.
7. J. W. Shin, N. S. Kim, "Perceptual reinforcement of speech signal based on partial specific loudness," IEEE Signal Processing Letters, to appear.
8. R. J. McAualy and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-28, 137-145, Apr. 1980.
9. Russell J. Niederjohn and James H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by highpass filtering followed by rapid amplitude compression," in Proc. of ICASSP, Aug. 1976, 24, 277-282.
10. ITU-T P.800, Methods for Subjective Determination of Transmission Quality, Aug. 1996.

저자 약력

•최 재 훈 (Jae-Hun Choi)

2007년 2월 : 인하대학교 전자공학과 학사
2007년 1월~2008년 2월 : 삼성전자 정보통신 총괄 연구원
2008년 3월~현재 : 인하대학교 전자공학부 석사과정

•장 준 혁 (Joon-Hyuk Chang)

1998년 2월 : 경북대학교 전자공학과 학사
2000년 2월 : 서울대학교 전기공학부 석사
2004년 2월 : 서울대학교 전기컴퓨터공학부 박사
2000년 3월~2005년 4월 : 마젠다스 연구소장
2004년 5월~2005년 4월 : 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월~2005년 8월 : 한국과학기술연구원 (KIST) 연구원
2005년 9월~현재 : 인하대학교 전자공학부 조교수