논문 2009-46SP-1-12

음성강화를 위한 시간 및 주파수 도메인의 분산정규화 기반 잡음예측 및 저감방법

(Nose Estimation and Suppression methods based on Normalized Variance in Time-Frequency for Speech Enhancement)

이 수 정*, 김 순 협**

(Soojeong Lee and Soonhyob Kim)

요 약

잡음예측 및 저감방법은 음성통신과 인식분야의 중요한 핵심기술이다. 본 논문에서는 다양한 잡음환경에 적용할 수 있는 새로운 잡음예측 및 저감 방법을 제안한다. 제안된 알고리즘은 시간 및 주파수영역의 noisy power spectrum 의 분산과 그 값의 정규화 ratio를 기반으로 한다. 제안한 방법은 다양한 잡음환경에서 잘 동작 할 수 있도록 적응추적 임계값을 사용하며, 이 임계값은 음성왜곡과 잔여잡음 사이의 trade-off를 제어한다. 새로운 알고리즘의 성능은 다양한 잡음환경에서 ITU-T P.835 (SIG) and segment (SNR) 의해 평가하여 기존의 방법에 비해 향상된 결과를 나타냈다.

Abstract

Noise estimation and suppression are a crucial factor of many speech communication and recognition systems. In this paper, proposed algorithm is based on the ratio of variance normalized of noisy power spectrum in time-frequency domain . Our proposed algorithm tracks the threshold and controls the trade-off between residual noise and distortion. This algorithm is evaluated by the ITU-T P.835 signal distortion (SIG) and segment signal to noise ratio (SNR), and is superior to the conventional methods.

Keywords : Noise estimation, Speech enhancement, Noise suppression.

I. 서 론

잡음예측 및 저감방법은 다양한 음성통신기기분야의 중요한 핵심요소기술이다. 잡음에 오염된 음성신호의 음질과 명료도 향상을 통하여 음성통신시스템의 성능을 향상시킬 수 있다. 잡음저감방법의 성능은 잡음구간 예측의 정교함 및 음성왜곡과 잔여잡음제거 사이의 trade-off를 제어하는 특성에 의존한다[1]. 따라서 음성왜곡 없이 잡음을 효과적으로 제거하고, 음질을 향상시키는 것은 매우중요하며 어려운 일이다.

현재 가장 잘 알려진 잡음제거 기술 중 spectral subtraction(SS)[1] 있다. 이 방법은 간단하면서도 음성왜곡 및 잔여잡음을 제어할 수 있어 많은 분야에 응용되고 있다. 그러나 spectral subtraction(SS)[1]의 가장 큰 문제점인 "musical noise"로 인해 시스템의 성능을 저하시키는 것이다[2]. 또 다른 방법으로 minimum mean square error (MMSE)[3]과 wiener[4] 방법이 사용되고 있지만, 이 방법들 역시 잔여잡음으로 인해 만족스러운 성능을 보여주지 못하였다. 이러한 기존 방법의 공통적인 특징은 음성의 처음과 끝부분을 묵음으로 추

* 정회원, 성균관대학교, BK21 정보기술사업단
(BK 21 post-doc, Sungkunkwan University)
** 정회원, 광운대학교, 컴퓨터공학과 교수.
(Professor, Computer Eng, Kwangwoon University)
※ 이 연구는 광운대학교 2008년 이 러닝 HCI 기술연구센터 지원으로 수행 되었습니다 .
접수일자: 2008년4월30일, 수정완료일: 2008년12월29일

정하는 음성검출방법 voice active detector (VAD)[5] 를 사용하지만, 실제 잡음환경과 낮은 signal-to-noise ratio (SNR) 에서는 시스템의 효율을 급격히 감소시킨다. 최근 VAD 아닌 비정상 잡음환경에 적용할 수 있는 알고리즘들이 활발히 연구·발표되고 있다. 대표적인 방법들을 살펴보면 minimum statistic (MS)[6], minima controlled recursive average (MCRA)[7], improve MCRA[8], 또한 MMSE를 개선시킨 a noncausal a priori (NCAP)[9] 방법이 발표되었고, 2006년에는 MCRA2[10] 방법이 발표되었다. 그러나 이들 방법 역시 잡음레벨이 급격히 변하는 비 정상잡음환경에서 잡음의 최소치 검색을 위한 응답지연시간으로 때문에 알고리즘의 성능이 현저히 저하된다. 또 다른 문제점으로는 잡음예측 알고리즘을 이용하여 잡음을 예측하고, 기존의 잡음제거 알고리즘인 SS 또는 MMSE 와 결합하여 추정된 잡음을 제거하기 때문에 알고리즘이 복잡해지는 단점을 가진다.

본 논문에서는 이러한 단점들을 최소화하고, 잡음예측부분과 잡음 제거 후 생기는 인공잔여잡음 (musical noise) 을 제거 하는 새로운 방법을 제안한다. 제안된 알고리즘은 시간 및 주파수 축에서 평균과 분산을 기반으로, 분산의 정규화 값을 구하고, 각 구간별 분산과 정규화 값의 비율을 계산한다. 비정상 및 정상 잡음환경에서 이 비율을 음성구간 및 비 음성 구간을 구분하는 방법으로 사용하였다. 특히, 음성이 포함된 구간에서 이 비율은 비 음성구간에 비해 매우 큰 값을 가진다. 특히, 제안한 방법은 비정상 잡음환경에서도 잘 동작할 수 있도록 잡음을 추적하는 방법으로 적응 임계 치 기법을 사용하였다[11]. 이 방법으로 기존알고리즘의 가장 큰 문제점인 음성왜곡과 잔여잡음간의 균형문제를 최소화 하였고, 잔여잡음제거에는 수정된 저감 함수를 이용하여 기존알고리즘에 비해 우수한 성능을 나타내었다. 음성왜곡과 잔여잡음간의 trade off 를 조절하기 위해 시간 및 주파수 축에서 δ_t 와 δ_f의 적절한 변수 값을 사용한다. 개선된 noisy power spectrum 과 저감 함수를 사용하여 향상된 speech power spectrum을 계산하였고, 제안한 알고리즘 평가에는 segmental SNR 과 ITU-T [12] 의 방법을 사용하였다. 사용된 잡음조건은 백색잡음, 버블잡음 과 자동차 잡음을 사용하였으며, 실험결과 제안한 방법이 기존방법들에 비해 우수한 성능을 보여주었다. 특히, 백색잡음제거에 탁월한 성능향상을 나타냈다.

본 논문의 구성은 다음과 같다. Ⅱ장에는 신호모델을 소개하고, Ⅲ장에서는 제안한 알고리즘의 잡음예측 및 저감방법을 다루었고, Ⅳ장에서는 실험결과를 보여주고 있다. 마지막으로 Ⅴ장에서는 제안한 방법의 결론 및 향후 과제를 언급하였다.

Ⅱ. 신호 모델

본 논문에서는 음성신호와 잡음신호간의 무상관으로 가정한다. 수식 (1) 은 잡음에 오염된 음성신호를 나타낸다. 수식 (1) 에서 $s(n)$은 음성신호, $d(n)$ 잡음신호를 나타낸다. 잡음에 오염된 음성신호를 Hamming window 함수를 이용하여 중첩된 구간들로 나누고, 단구간 Fourier 변환(STFT)[13]을 사용하여 각각의 구간들을 시간 및 주파수 축으로 나타낸다. 여기서 $k \rightarrow k(1 \le k \le K)$는 주파수 인덱스이며, $l \rightarrow l(1 \le l \le L)$은 프레임(즉, 시간 축 인덱스) 이다. 아래 수식 (2)는 잡음이 오염된 음성스펙트럼을 좀 더 구체적으로 표현하고 있다.

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

$$|X(k,l)| = \sum_{n=0}^{N-1} x(n+lL)w(n)e^{-j(\frac{2\pi}{N})nk} \tag{2}$$

여기서 w 는 윈도우 함수를 나타내고, N 는 윈도우 크기, L 프레임 step 을 나타낸다. 아래 수식(3)은 잡음에 오염된 음성신호의 power spectrum을 나타낸다.

$$|X(k,l)|^2 = |S(k,l)|^2 + |D(k,l)|^2 \tag{3}$$

여기서 $|S(k,l)|^2$는 음성신호의 파워스펙트럼 이며, $|D(k,l)|^2$는 잡음신호의 파워스펙트럼이다.

Ⅲ. 잡음예측 및 저감 알고리즘

제안된 잡음예측 및 저감알고리즘은 시간 및 주파수 축의 잡음에 오염된 음성 power spectrum 분산에 기반한다.

$$\mu_t(l) = \frac{1}{K}\sum_{k=1}^{K}|X(k,l)|^2,$$

$$\mu_f(k) = \frac{1}{L}\sum_{l=1}^{L}|X(k,l)|^2 \tag{4}$$

여기서 $\mu_t(l)$는 각 주파수 구간의 잡음에 오염된 음성파워스펙트럼의 평균을 나타내며, $\mu_f(k)$ 는 각 프레임의 잡음에 오염된 음성파워스펙트럼의 평균을 나타낸다.

$$v_t(l)=\frac{1}{K}\sum_{k=1}^{K}(|X(k,l)|^2-\mu_t(l))^2 \tag{5}$$

$$v_f(k)=\frac{1}{L}\sum_{l=1}^{L}(|X(k,l)|^2-\mu_f(k))^2 \tag{6}$$

여기서, 수식 (5)와 (6)의 $v_t(l)$과 $v_f(k)$ 는 시간 및 주파수 축의 오염된 잡음 파워 스펙트럼의 분산이며, 수식(7)은 수식(5) 와 (6)의 정규화 값으로, 시간-주파수 축의 잡음파워 추정 값으로 사용한다. 다음, 수식 (8)은 시간과 주파수 축에서 noisy power spectrum 의 분산과 정규화 값의 비율로 결정된다.

$$\widehat{\sigma_t^2}=\frac{1}{L}\sum_{l=1}^{L}v_t(l),\ \widehat{\sigma_f^2}=\frac{1}{K}\sum_{k=1}^{K}v_f(k) \tag{7}$$

$$\gamma_t(l)=\frac{v_t(l)}{\widehat{\sigma_t^2}},\ \gamma_f(k)=\frac{v_f(k)}{\widehat{\sigma_f^2}} \tag{8}$$

수식 (8)의 값 (즉, 분산과 정규화 비율)이 적응 임계 값 (threshold)보다 큰 값을 가질 경우 음성구간으로, 작은 값을 가질 경우 비 음성구간으로 구분한다. 일반적으로, 음성이 포함된 구간은 발성된 음성스펙트럼이 잡음스펙트럼에 영향을 주어 각 시간 및 주파수 축의 평균 및 분산 값을 변화시킨다. 그러므로 음성이 포함된 구간은 음성이 포함되지 않은 구간에 비해 매우 큰 값을 가진다. 상대적으로 음성이 포함되지 않은 구간 (잡음구간)으로 추정된 부분은 매우 작은 값을 (0 과 가까운 값) 나타낸다. 이런 특성으로 음성구간과 비 음성구간을 구분한다.

1. 시간 축에서 적응 임계 치를 이용한 잡음예측 방법

제안한 알고리즘은 적응알고리즘을 이용하여 잡음임계 값을 추적하고, 음성왜곡과 잔여잡음의 trade-off을 조절 한다.

$$\xi_t(1)=\gamma_t(1)+\delta_t \tag{9}$$

$$\xi_{tL}=\xi_t(1)\cdot\delta_{tL} \tag{10}$$

$$\xi_{tU}=\xi_t(1)\cdot\delta_{tU} \tag{11}$$

$$\alpha_t(t)=\xi_t(l-1)-\gamma_t(l-1) \tag{12}$$

$$\begin{aligned}&IF\ \ \alpha_t(l)>\delta_t,\\&\quad\xi_t(l)=\xi_t(l-1)\cdot(1-\zeta_a)+\xi_{tL}\cdot\zeta_a\\&ELSE IF\ \ \zeta_z\le\alpha_t(l)\le\delta_t,\\&\quad\xi_t(l)=\xi_t(l-1)\cdot(1-\zeta_b)+\xi_{tU}\cdot\zeta_b\\&ELSE\\&\quad\xi_t(l)=\xi_t(l-1)\cdot(1-\zeta_a)+\xi_{tU}\cdot\zeta_a\end{aligned} \tag{13}$$

수식(9) 에서 제어변수 δ_t 정의하고, 수식 (8) 의 초기치를 이용해 적응 임계 초기 값을 구한다. $\xi_t(l)$는 이전구간의 $\xi_t(l-1)$ 사용하는 적응 임계 값 이며, ξ_{tL} 는 적응 임계 치 하한 값으로 적응 임계 치의 초기 값 수식(9) 와 실험을 통해 얻어진 상수 0.5를 곱하여 임계 치의 하한 값으로 사용한다. 일반적으로 적응 임계 치 초기 값 은 묵음구간 (즉, 비 음성)으로 잡음만 존재한다고 가정한다. 만약 특정구간이 적응 임계 치의 하한 값 보다 작은 값을 가질 경우 해당구간은 비 음성구간으로 가정한다. 또한 ξ_{tU} 는 적응임계 값의 상한 값으로 적응 임계 치 초기 값에 상한 경계 상수 2.0을 곱하여 표현한다. 즉, 적응 임계 초기 값의 2배 이상의 값을 가진 구간은 음성구간으로 가정한다. 수식(11)의 $\alpha_t(l)$ 는 이전 구간의 적응 임계 치 값 $\xi_t(l-1)$ 과 수식 8의 분산과 그 값의 정규화 비율 $\gamma_t(l-1)$ 이용하여 구할 수 있다. 이 값을 적응 임계 치를 추적하는 추적 변수로 사용한다. 적응 임계 치 $\xi_t(l)$는 추적변수 $\alpha_t(l)$의 값에 따라 변화한다.

본 논문에서 제안한 방법은 $\xi_t(l)$ 적응 임계 치를 구할 때 기존의 Minimum mean square error (MMSE) 나 Spectral subtraction (SS) 에서 사용한 비 음성 구간에서 잡음의 예측 치를 추정하는 방법이 아닌 이전구간 $\xi_t(l-1)$ 적응 임계 치를 이용한다. 이 방법을 사용할 경우 기존 방법의 단점인 잡음 예측 치를 구하는 동안 발생하는 응답시간지연 문제점을 해결 할 수 있다. 만약, 추적변수 $\alpha_t(l)$의 값이 제어변수 δ_t 보다 큰 값을 가질 경우(즉, 이전구간 적응 임계 치 와 수식 (8)의 정규화 비율의 값의), 하한 임계 치에 가중치 상수 0.8을 곱하고, 이전구간의 적응 임계 치에 0.2 을 곱하여

이전구간 적응 임계 치에 비해 감소된 값을 현재구간의 적응 임계 치 값으로 사용한다. 결론적으로 적응 임계 치는 상한 과 하한 임계 치 사이에서 정규화 비율을 추적하는 것이다. 또한, $\alpha_t(l)$ 값이 0 근사한 양수 값을 가질 경우 이전 임계 치에 가중치 상수 0.2에서 0.6 으로 증가시켜 현재 적응 임계 치에 적용한다. 다음, $\alpha_t(l)$ 값이 0 보다 작은 음수 값을 가질 경우, 상한 임계 치와 가중상수 0.8 을 사용해 적응 임계 치를 증가시킨다. 위에서 사용한 가중치 상수를 반복 된 실험을 통하여 얻어진 값 들이다. 그림 1과 2에서 제어변수 δ_t에

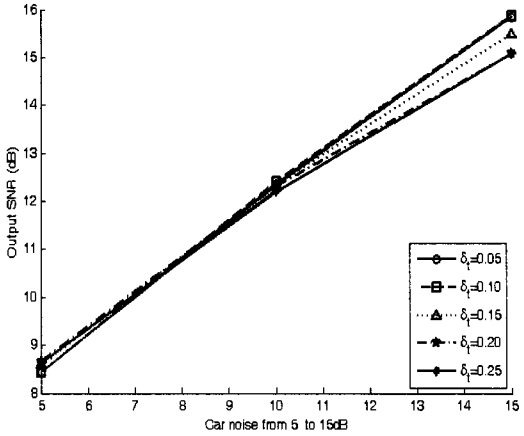

그림 1. Car noise SNR 5dB 에서 15dB 환경에서 제어변수 δ_t 의 출력 SNR (dB) 에 대한 효과

Fig. 1. Effect of various δ_t values on SNR gains with car noise environments in the range SNR 5(dB)- 15(dB).

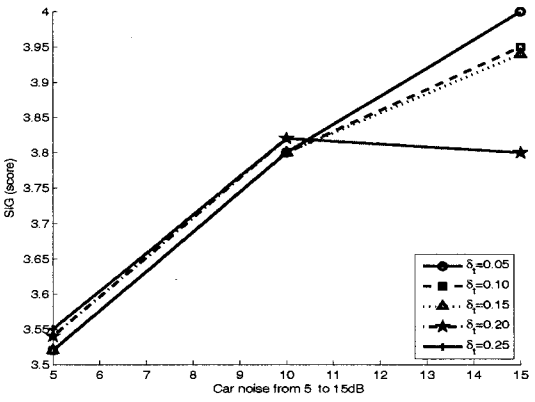

그림 2 Babble and car noise SNR 5dB 에서 15dB 환경에서 제어변수 δ_t의 출력 SIG (score) 에 대한 효과

Fig. 2. Effect of various δ_t values on SIG score with babble and car noise environments in the range SNR 5 (dB)- 15(dB).

의해 음질왜곡과 잔여잡음 간의 trade-off한 특성을 보여주고 있다. 실험결과 5 dB 이하의 SNR 일 경우 제어변수 δ_t값을 증가시켜 출력 SNR 을 높일 수 있고, 15 dB 이상에서는 제어변수 δ_t 값을 감소시켜 출력 SNR 을 증가시킬 수 있다. 여기서 제어변수 δ_t의 초기 값으로 상수 0.1을 설정하였지만, 입력 SNR 이 변하는 비정상 잡음 환경에서는 고정된 제어변수 δ_t을 사용할 경우 비효율적인 특성을 확인 하였다.

예를 들어 그림 1과 2에서 입력 SNR 5dB 에서 제어변수 δ_t의 값을 0.25 로 설정한 경우 출력 SNR 을 증가시킬 수 있지만, SNR 15dB 에서는 음성신호가 왜곡된다. 실제 시스템의 경우 음성왜곡이 시스템에 치명적인 결과를 초래할 수 있다. 그림 2는 제어변수 δ_t의 값에 따라 음성신호의 왜곡(SIG)을 보여준다. 실험결과 car noise 5dB 이하의 SNR 에서는 제어변수 δ_t을 증가시키고, 15dB 이상의 SNR 에서는 제어변수 δ_t을 감소시키는 것이 효과적이다. 결론적으로 주요 잡음에 대한 특성에 따라 제어변수 δ_t을 사용하여 음성신호의 왜곡과 잔여잡음 제거를 효과적으로 제어할 수 있다. 서론에서 언급한 MS 알고리즘의 잡음추정 방법은 한정된 구간의 특정 윈도우에 대한 오염된 음성신호의 power spectrum의 최소치를 통하여 얻을 수 있다. 그림 3에서, babble noise SNR 10dB 다음

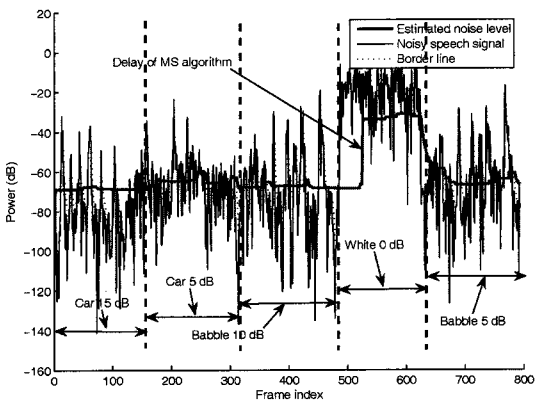

그림 3. 비정상 잡음환경 car noise (15dB and 5dB), babble noise (10dB), white Gaussian noise 0(dB) and babble noise 5(dB)에서 MS 알고리즘의 오염된 음성 파워스펙트럼과 잡음추정

Fig. 3. Noisy power spectrum and noise estimate of MS method for car noise (15dB and 5dB), babble noise (10dB), white Gaussian noise 0dB and babble noise 5dB in a nonstationary at f=625 Hz.

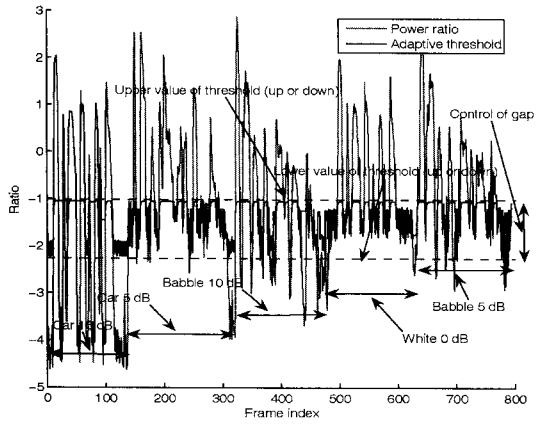

그림 4. 비정상 잡음환경 car noise (15dB and 5dB), babble noise (10dB), white Gaussian noise 0(dB) and babble noise 5(dB)에서 제안한 알고리즘의 오염된 음성 파워스펙트럼의 적응 임계치 추정

Fig. 4. Adaptive thresholds estimation on the time index for car noise (15dB and 5dB), babble noise 10(dB), white noise 0(dB) and babble noise 5(dB) in a nonstationary noisy environments.

white noise SNR 0dB 로 잡음레벨이 급격히 증가하는 구간에서 일정 구간만큼 응답시간지연이 생기는 현상을 볼 수 있다.

여기서, 그림 4의 제안한 방법은 잡음레벨이 급격히 증가할 경우에도 이전 구간의 적응 임계 치를 현 구간 임계 치로 사용하므로 응답시간지연이 나타나지 않는다. 만약, 제안한 방법에서 음성왜곡이 나타날 경우 제어변수 δ_t를 감소시켜 문제를 해결하고, 잔여잡음이 나타날 경우 제어변수 δ_t값을 증가시켜 해결한다. 특히 제안한 방법이 잡음레벨이 높은 car 5 dB와 white noise 0 dB의 비정상 잡음환경(잡음레벨이 시간에 따라 변하는)에서 기존방법에 비해 추적성능이 우수한 결과를 보인다.

아래 수식 (14) 에서 앞에서 구한 적응 임계 치를 수식 (8)의 분산의 정규화 비율과 비교하여 음성과 비 음성을 구분하는 방법을 나타낸다.

$$IF\ \ \gamma_t(l) \le \xi_t(l), \tag{14}$$

$$ELSE$$

$$\widehat{D_n}(k,l) = \widehat{D_r^2}(k,1)$$

$$\widehat{D_r^2}(k,1) = \frac{1}{l}\sum_{m=1}^{l}\left(\frac{1}{K}\sum_{k=1}^{K}\widehat{D_n^2}(k,l)\right)$$

$$\widehat{D_n^2}(k,l) = |X(k,l)|^2,$$

$$G_{up}(k,l) = G(k,l) \cdot 0.001$$

$$G_u(k,l) = G(k,l) \cdot 0.99$$

여기서 저감 상수를 0.001로 초기화하고, 저감 함수 $G(k,l)$는 1.0으로 초기화 한다. 위의 알고리즘을 설명하면, 수식(14)의 $\gamma_t(l)$ 을 적응 임계 치 $\xi_t(l)$과 비교하여, 값 $\gamma_t(l)$이 $\xi_t(l)$보다 작은 값을 나타내면 해당 구간을 비 음성 구간으로, 그렇지 않을 경우 음성 구간으로 가정한다. 만약 해당구간이 비 음성 구간일 경우 오염된 음성 power spectrum $|X(k,l)|^2$의 해당구간을 $\widehat{D_n^2}(k,l)$로 할당한다. 여기서 할당된 $\widehat{D_n^2}(k,l)$는 잡음 power spectrum 으로 추정하고, $\widehat{D_r^2}(k,1)$는 비 음성구간의 누적 치의 평균값으로 나타낸다. 여기서 $\widehat{D_r^2}(k,1)$ 값은 음성구간의 잔여 잡음 값으로 추정하여, 잔여잡음을 제거하는데 사용한다. 위 수식(14)의 $\widehat{D_r^2}(k,1)$는 $\widehat{D_n^2}(k,l)$ 로 할당된다. 위 수식에서 비 음성 구간은 저감 함수 $G(k,l)$에 0.001을 곱하며, 음성구간에서는 0.99를 곱하여 개선된 저감 함수 $G_{up}(k,l)$을 구한다. 마찬가지로, 저감 함수도 잔여잡음을 제거하는 용도로 사용한다.

$$|X_{up}(k,l)|^2 = |X(k,l)|^2 - \widehat{D_n^2}(k,l) \tag{20}$$

$$|X_{up}(k,l)|^2 = MAX(|X_{up}(k,l)|^2, 0.001) \tag{21}$$

수식 (20)에서 비 음성구간의 잡음과 음성구간의 잔여잡음을 제거한다. 또한, 수식(21)은 수식 (20)의 연산결과의 음수를 제거하는 용도로 사용된다.

2. 주파수 축에서 적응 임계 치를 이용한 잡음예측방법

시간 도메인에서 사용한 방법을 동일한 방법을 주파수 도메인에 적용한다.

$$\xi_f(1) = \gamma_f(1) + \delta_f \tag{22}$$

$$\xi_{fL} = \xi_f(1) \cdot \delta_{fL} \tag{23}$$

$$\xi_{fU} = \xi_f(1) \cdot \delta_{fU} \tag{24}$$

$$\alpha_f(k) = \xi_f(k-1) - \gamma_f(k-1), \tag{25}$$

$$IF\ \alpha_f(k) \geq \delta_f\ ,$$
$$\xi_f = \xi_f(k-1) \bullet (1-\eta_a) + \xi_{fL} \bullet \eta_a \tag{26}$$

$$ELSEIF\ \eta_z \leq \alpha_f(k) \geq \delta_f$$
$$\xi_f = \xi_f(k-1) \bullet (1-\eta_a) + \xi_{fU} \bullet \eta_b$$
$$ELSE$$
$$\xi_f = \xi_f(k-1) \bullet (1-\eta_a) + \xi_{fU} \bullet \eta_a$$

그림 5는 주파수 도메인의 제어변수 δ_f 에 의한 적응 임계 치 $\xi_f(k)$ 를 보여주고 있다. 그림 6과 7은 제어변수 δ_f에 의한 SNR 게인과 SIG를 나타내고 있다. 실험 결과, 제어변수 δ_f에 대한 최적의 값은 0.001로 나타났다. 시간 축과는 다르게 입력 SNR (dB)에 따라 출력 SNR (dB) 이 변하는 것이 아닌, δ_f가 증가 할수록 출력 SNR 과 SIG 가 감소한다는 결과를 확인하였다. 이러한 특성은 그림 5의 (붉은 점선)에서 3500 Hz 이상의 모든 잡음들이 제거되며, 동시에 음성신호의 자음들도 제거되어 음성신호의 명료함이 저하되는 현상으로 나타났다. 이와 반대로 δ_f을 감소시키면 SNR 과 SIG 가 증가하며 음성신호의 명료함도 증가하지만, 3500 Hz 이상의 잔여 잡음으로 인한 불쾌한 잡음소리를 확인하였다. 향후, 이러한 문제점을 해결하여 음성신호의 왜곡을 최소화 하고자 한다. 또한, 시간 도메인에서는 출력 SNR 이 SIG와 trade-off 한 기존의 이론적 특성을 확인하였지만, 주파수 도메인에서 출력 SNR 이 SIG와 trade-off 한 특성이 아닌 SNR이 증가 하면, SIG도 함께 증가하는 특성으로 나타났다. 여기서 SNR은 잡음제거의 특성을 표현하고, SIG는 음성왜곡의 수치를 나타낸다.

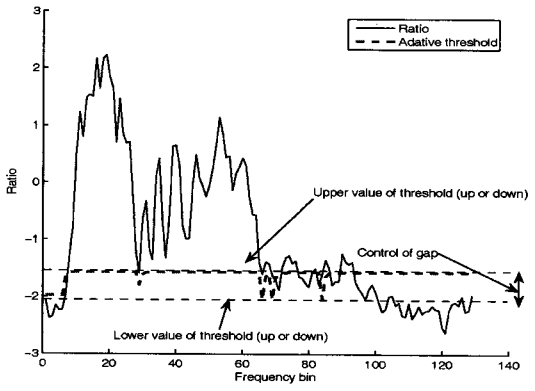

그림 5. 비정상 잡음환경의 주파수 도메인에서 적응임계치 추정

Fig. 5. Adaptive thresholds estimation on the frequency bin index in a nonstationary noisy environment.

아래 수식 (28)는 주파수 도메인에서 음성 및 비 음성 대역의 잔여잡음을 제거하기 위해 시간도메인에서 구한 개선된 저감 함수에 저감 상수 0.001을 곱하여 시간 및 주파수 도메인에서 새롭게 갱신된 저감 함수를 구한다.

$$IF\ \gamma_f(k) \leq \xi_f(k), \tag{27}$$

$$G_{mo}(k,l) = G_{up}(k,l) \bullet 0.001 \tag{28}$$

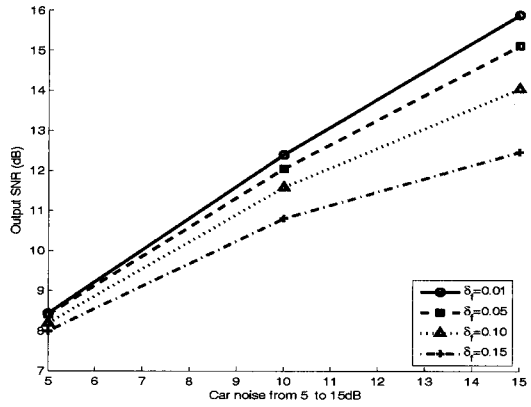

그림 6. Car noise SNR 5dB 에서 15dB 환경에서 제어변수 δ_f의 출력 SNR (dB) 에 대한 효과

Fig. 6. Effect of various δ_f values on SNR (dB) with car noise environments in the range SNR 5 (dB)- 15 (dB).

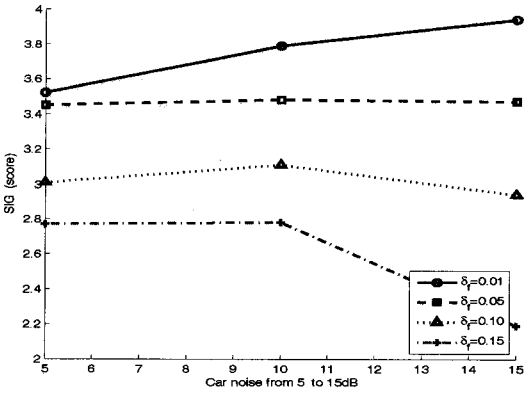

그림 7. Car noise SNR 5dB 에서 15dB 환경에서 제어변수 δ_f의 출력 SIG (score) 에 대한 효과

Fig. 7. Effect of various δ_f values on SIG (score) with car noise environments in the range SNR 5 (dB)- 15 (dB).

$$G(k,l) = G_{mo}(k,l) \tag{29}$$

ELSE

$$G_{mo}(k,l) = G_{up}(k,l) \cdot 0.99 \tag{30}$$

$$G(k,l) = G_{mo}(k,l) \tag{31}$$

마지막으로, 개선된 음성 power spectrum $|\hat{S}(k,l)|^2$ 은 $G(k,l)$ 저감 함수와 개선된 noisy power spectrum $|X_{up}(k,l)|^2$ 을 가지고 아래와 같이 나타낼 수 있다.

$$|\hat{S}(k,l)|^2 = G(k,l) \cdot |X_{up}(k,l)|^2 \tag{32}$$

V. 실험 결과

본 논문에서 제안된 noisy power spectrum 의 표준편차와 그 값의 평균의 정규화 ratio 방법이 적용된 음성향상 알고리즘의 성능평가를 위해 white, babble, and car noise 환경에서 객관적 테스트를 수행하였다. 실험에 사용된 DB 는 남성과 여성 각각 3명이 발성한 NOIZEUS[12]을 사용하였고, 음성신호를 8kHz로 sampled, short time Fourier transform (STFT)를 사용하여 50% overlapping Hamming windows 256 samples을 이용하였다. 표 1은 기존방법 및 제안한 방법의 segmental SNR 을 보여주고 있다. Segmental SNR 결과로부터 제안한 방법이 기존의 방법들보다 향상된 결과를 나타냈고, 특히 백색잡음 환경에서는 우수한 결과를 보였다.

표 2는 다양한 잡음환경에서 SIG 결과를 보여준다.

표 1. White, babble and car noise 환경에서 segmantal SNR (dB)

Table 1. Segmantal SNR (dB) at white, babble and car noisy environment.

method	SNR (dB)	white noise	babble noise	car noise
SSMS	5	7.29	6.33	5.66
	10	11.62	10.68	10.99
	15	15.66	15.24	15.10
WIENERWT	5	10.30	8.90	6.14
	10	14.48	12.25	9.90
	15	17.85	16.24	15.19
PROPOSED	5	10.83	8.43	6.39
	10	14.75	12.40	11.65
	15	18.28	15.90	15.74

표 2. 제안한 음성 향상 알고리즘과 (PROPOSED) 과 기존 방법 (SSMS) 와 (WIENERWT)[12] 방법의 SIG 수치 비교

Table 2. SIG result for the proposed speech enhancement method and conventional methods, 5= no degradation, 4=little degradation, 3=somewhat degradation, 2=fairly degradation, and 1=very degraded.

method	SIG (score)	white noise	babble noise	car noise
SSMS	5	1.65	2.69	3.22
	10	2.28	3.75	3.96
	15	2.96	3.90	3.13
WIENERWT	5	2.43	2.47	2.45
	10	3.33	3.79	3.28
	15	3.83	3.94	3.63
PROPOSED	5	3.07	3.52	3.13
	10	4.42	3.79	3.62
	15	4.64	3.94	3.66

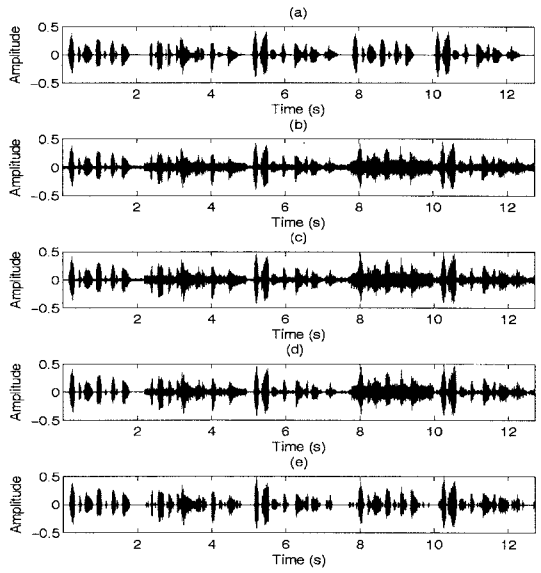

그림 8. Car noise 15 dB, car noise 5 dB, babble noise 10 dB, white noise 0 dB and babble noise 5 dB SNR 의 비정상 잡음환경에서 음성강화 알고리즘의 시간 축 결과. (a) 음성신호, (b) 잡음에 오염된 음성신호, (c) SSMS를 사용한 음성강화, (d) WIENERWT 를 사용한 음성강화 (e) 제안한 방법을 이용한 음성강화.

Fig. 8. Time domain results of speech enhancement for car noise at 15 dB, car noise at 5 dB, babble noise at 10 dB, white noise 0 dB, and babble noise 5 dB SNR in a nonstationary environment. (a) Original speech; (b) Noisy speech; (c) Enhancement speech using SSMS; (d) Enhancement speech using WIENERWT; (e) Enhancement speech using Proposed.

대부분의 잡음환경에서 제안한 방법이 기존에 방법에 비해 우수한 결과를 나타내었다.

VI. 결 론

본 논문에서는 시간-주파수 도메인 기반의 분산의 정규화 기법을 제안하였다. 제안된 방법에서 제어변수를 이용하여 음성왜곡과 잔여잡음을 제어하였고, 적응 임계 치를 이용하여 음성 및 비 음성구간을 구분하였다. 새롭게 제안된 결과는 기존의 방법에 비해 향상된 경과를 나타냈고, 특히 백색잡음 환경에서는 우수한 성능향상을 보여주었다.

참 고 문 헌

[1] M. Bhatnagar, "*A modified spectral subtraction method combined with perceptual weighting for speech enhancement,*" Master's thesis, University of Texas at Dallas, pp.1-10, 2003.

[2] S. F. Boll, "*Suppression of acoustic noise in speech using spectral subtraction,*" IEEE Trans. Acoust. Speech Signal Processing, 27, (2), pp. 113-120, 1979.

[3] Y. Ephraim and D. Malah, "*Speech enhancement using a minimum mean-square errror short-time spectral amplitude estimator,*" IEEE Trans. Acoust. Speech Signal Processing, 32(6), pp. 1109-1121, 1984.

[4] O. Cappe, "*Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,*" IEEE Trans. Speech Audio Processing, 2 (2), pp. 346-349, 1994.

[5] Y. Hu, "*Subspace and multitaper methods for speech enhancement,*" Ph.D. dissertation. University of Texas at Dallas, pp. 1-15, 2003.

[6] R. Martin, "*Noise power spectral density estimation based on optimal smoothing and minimum statistics,*" IEEE Trans. Speech Audio Processing, 9 (5), pp. 504-512, 2001.

[7] Y. Hu and P. Loizou, "*Speech enhancement based on wavelet thresholding the multitaper spectrum,*" IEEE Trans. Speech Audio Processing, 12 (1), pp. 59-67, 2003.

[8] I. Cohen, "*Noise spectrum in adverse environments: improved minima controlled recursive averaging,*" IEEE Trans. Speech Audio Processing, 11(5), pp. 466-475, 2003.

[9] I. Cohen, "*Speech enhancement using a noncausal a priori SNR estimator,*" IEEE Signal Processing Letters, 11 (9), pp. 725-728, 2004.

[10] R. Sundarrajan and C. L. Philipos, "*A noise-estimation algorithm for highly nonstationary noisy environments,*" Speech Communication, 48, pp. 220-231, 2006.

[11] ITU-T, "*Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*", ITU-T Recommendation P.835, 2003.

[12] C. L. Philipos, "*Speech Enhancement (Theory and Practice,*" 1st edition. CRC Press, Boca Raton, FL, 2007.

저 자 소 개

이 수 정(정회원)

1997년 한국방송통신대학교 전자계산학과 학사.

2000년 광운대학교 공학석사.

2008년 광운대학교 공학박사.

2008년 현재 성균관대학교 BK21 정보기술 사업단 박사 후 연구원

<주관심분야 : 음성신호처리, 잡음예측 및 제거, 적응신호처리>

김 순 협(정회원)

1983년 연세대학교 전자공학과 공학박사.

1998년~2000년 한국음향학회 회장.

2001년~현재 한국음향학회 명예회장

2001년~현재 한국음성정보처리 산업협의회 부위원장.

2002년~현재 한국음성정보처리 기술협력위원회 위원장.

1979년~현재 광운대학교 전자정보공과대학 컴퓨터공학과 교수.

<주관심분야 : 음성인식, 신호처리, HCI>