# 잡음 신호의 지각 패턴 제어를 통한 음질 개선 알고리즘 개발에 관한 연구

김헌중, 차형태
숭실대학교 전자공학과

# The Study for Noisy Speech Improvement with Noise Perception Pattern Suppression

Hunjoong Kim, Hyungtai Cha
Dept. Of Electronics Engineering, Soongsil Univ.
Email : hjkim@mms.ssu.ac.kr

## 요약

본 논문에서는 사람의 청각 모델을 기반으로 잡음에 의해 손상된 음성 신호로부터 잡음 신호의 마스킹 특성과 신호에너지의 지각(知覺)을 나타내는 임계대역(critical band)에서의 잡음 에너지에 대한 지각 패턴인 noise excitation pattern을 이용한 잡음 에너지 차감과 잡음 추정 오차에 의한 변형된 음성신호 내의 순음(tonal) 성분과 비순음(non-tonal)성분의 보정을 통해 효과적인 음성 품질의 개선을 위한 연구를 하였다.

## 1. 서론

잡음이 있는 환경에서의 음성 향상의 문제는 음성 인식, 다양한 통신환경, 음성 신호의 복원 및 부호화에서의 전처리 과정등에서 매우 중요한 역할 을 한다. 이와 같이 다양한 잡음 환경에서 음성의 품질을 개선 시키기 위한 방법에는 주파수 차감 기반의 방법[1][2], 소프트 디시젼(soft-decision) 필터링 방법[3], 최소평균제곱오차(MMSE)방법[4], 사람의 청각모델 특성을 이용한 방법[5][6]등 다양한 criteria을 통해 noise suppression rule을 적용하고 있다.

본 논문에서는 단일 채널 기반의 환경에서 음성 향상을 위해 잡음 신호에 대한 사람의 청각 시스템에서의 지각 패턴을 이용한 주파수 차감 기법 기반의 필터링을 통해 다양한 환경에서 음성 향상을 위한 연구를 수행하였다. 이것은 잡음에 오염된 음성 신호 내에서의 잡음 에너지 확산의 지각적인 에너지 간섭 효과를 perceptual *a posteriori* SNR을 통해 최소화 하고, 단일 채널 음성 향상 알고리즘에서 발생하게 되는 잡음 에너지 추정 오차에 의한 오염된 음성 신호 내의 변형된 음성 특성들을 보상 시켜 주는 방식이다.

## 2. Subtractive-Type Algorithm

임의의 음성 신호의 이산신호 표현을 $s(n)$ 이라고 하면, additive stationary background noise, $d(n)$ 에 의해 오염된 신호는 다음과 같이 표현 할 수 있다.

$$y(n) = s(n) + d(n) \qquad (0 \le n \le N-1) \tag{1}$$

이때, $F_y(\omega;t)$ 을 길이 $N$ 을 갖는 임의의 윈도우를 통해 구분되어지는 프레임 인덱스 $t$ 에서의 noisy 신호의 파워 스펙트럼이라고 하고, 음성신호와 잡음이 서로 비(非) 상관관계에 있다고 가정 한다면, 다음과 같이 표현 할 수 있다.

$$F_y(\omega;t) = F_s(\omega;t) + F_d(\omega;t) \tag{2}$$

일반적으로 단일 채널 음성 개선 알고리즘들의 경우 오직 하나의 입력 신호만 사용 가능하므로 오직 noisy 음성 신호만 이용 가능하다. 이러한 상황에서 음성 신호에 포함된 잡음의 특징 추출은 묵음 구간동안에 수행되어지게 되어, 개선된 음성 신호는 잡음신호 스펙트럼의 추정치를 통해 잡음신호의 에너지를 제어 함으로써 얻어 질 수 있다

$$\hat{F}_s(\omega;t) = F_y(\omega;t) - \hat{F}_d(\omega;t)$$

where, $\hat{F}_s(\omega;t)$ : the enhanced speech power spectrum
$F_y(\omega;t)$ : the noisy speech power spectrum
$\hat{F}_d(\omega;t)$ : the estimated speech power spectrum (3)

결국, 이러한 주파수 차감 형식의 알고리즘을 통해 얻을 수 있는 개선된 음성 신호는 묵음 기간 동안 추정한 잡음 스펙트럼과 noisy신호의 위상 정보의 조합에 의한 것이 된다.

그러나 이와 같은 주파수 차감 형식의 알고리즘은 주파수 차감 과정에서 발생하는 musical tone noise 와

residual noise reduction 처리 과정이 없으면 만족스러운 결과를 얻을 수 없게 되는데, 대부분의 알고리즘들은 이러한 문제들을 해결하기 위해 다양한 criteria 을 통해 차감 규칙을 정하고 있으며, 잡음 신호 에너지 제어에 있어, 다양한 환경의 변화에 적응할 수 있도록 파라미터를 통한 제어형태를 취하게 된다.

이때 추정 하고자 하는 개선된 음성 신호의 진폭 스펙트럼을 $\left|\hat{S}_s(w;t)\right|$라 한다면,

$$\hat{F}_s(\omega;t)=\left|\hat{S}_s(\omega;t)\right|^2 \tag{4}$$

다음은 주파수 차감 형식의 알고리즘의 파라미터를 통한 제어형태를 취한 일반적인 형태가 된다.

$$\hat{S}_s(\omega;t)=\left[\left|S_y(\omega;t)\right|^{\alpha}-k\left|\hat{S}_d(\omega;t)\right|^{\alpha}\right]^{1/\alpha}e^{j\theta_y(\omega;t)} \tag{5}$$

여기에서 지수 $\alpha$ 은 noise suppressing curve의 transition의 sharpness를 제어하는 요소이고, over-subtraction factor $k$ 는 일반적으로 SNR 근거한 noise suppresstion의 정도를 제어하게 된다.

이때 이러한 주파수 차감 형식의 알고리즘은 noisy 신호의 스펙트럼과 추정한 잡음 신호의 스펙트럼에 의존하는 임의의 필터 $H(\omega;t)$ 를 통해 다음과 같은 관계를 만족시키는 시스템으로 구성 할 수 가있다.

식(5)와 식(11)으로부터

$$\hat{F}_s(\omega;t)=H(\omega;t)\cdot F_y(\omega;t) \tag{6}$$

이때 $H(\omega;t)$ 는

$$\begin{aligned}H(\omega;t)&=H\left[SNR_{post}(\omega;t)\right]\\&=\left(1-k\left[\frac{\left|\hat{S}_d(\omega;t)\right|}{\left|S_y(\omega;t)\right|}\right]^{\alpha}\right)^{1/\alpha}\end{aligned} \tag{7}$$

이와 같이 $H(\omega;t)$ 는 잡음에 오염된 음성 신호 내에서 잡음에 의해 음성 신호가 얼마나 감쇄 하였는가를 나타내는 *a posteriori SNR*에 의해 감소시킬 잡음 에너지의 양을 결정 하게 된다[1][3][9].

## 2. Psychoacoustical Representation of Signals

이와 같이 주파수 차감 형식의 알고리즘들은 잡음 신호의 에너지 차감에 있어 고정된 파라미터들로는 변화하는 잡음 신호의 에너지 레벨과 특성들에 적응 하여 효과적으로 잡음 에너지를 제어하기 힘들 뿐만 아니라, 이러한 파라미터들을 다양한 환경에 맞게 최적화하는 것은 매우 어려운 일이 된다.

이러한 이유로 인해 일반적인 주파수 차감 형식의 음성 개선 알고리즘들은 다양한 환경에 대응할 수 있도록 자유 파라미터를 통해, 잡음 에너지 감소와 잔여 노이즈 에너지의 존재, 잡음 에너지 제어 과정에서 부수적으로 발생하게 되는 음성 신호의 원치 않는 왜곡 등의 사이에 존재하는 tradeoff에 대한 variation을 제공하게 되는데, 최근 인간의 청각적 지각특성을 이용하여 잡음 에너지 차감 파라미터를 마스킹 특성에 근거해서 적응시키거나[5], audible noise의 psychoacoustical noise *shaping[6]*을 통해 잡은 에너지 감소와 음성의 명료도 (intelligibility) 증가를 통해 많은 개선을 가져왔다.

사람의 청각 시스템에서의 주파수 변별력(frequency selectivity)을 나타내는 임계 대역(critical band)에서의 음성 신호의 파워 스펙트럼 $F_s(\omega;t)$ 에 대한 critical band intensity는 다음과 같이 계산 할 수 있다.

$$F_s^l(z;t)=a_0(z)\sum_{k=k_{lz}}^{k_{hz}}F_s(\omega;t)\qquad 0\le z\le Z-1$$

where, $z$ : critical band index (in *Bark*)

$k_{lz}, k_{hz}$ : the lower and upper bounds of teh critical band $z$

$Z$ : the total number of critical band

(8)

이 때, $a_0(z)$ 는 외이(outer ear)에서부터 중이(middle ear)까지의 다양한 transmission factor들에 의한 주파수에 따른 감쇄특성을 나타내는 attenuation factor 이다.

이러한 $F_s^l(z,t)$ 은 기저막(basilar membrane)에서의 신호의 에너지의 확산현상을 나타내는 확산 함수 (spreading function)를 통해 컨벌루션을 취함으로써 다음과 같이 기저막에서의 신호 에너지에 대한 주파수 응답을 나타내는 excitation level을 계산할 수 있다.

$$F_s^e(z;t)=\left[\sum_{v=0}^{z-1}\left[s_2(v;z-v)[F_s^l(v;t)]^{1+0.02(z-v)dz}\right]^{\delta/2}+\sum_{v=z}^{Z-1}\left[s_1(v-z)F_s^l(v;t)\right]^{\delta/2}\right]^{2/\delta} \tag{9}$$

여기에서, $s_1(v-z)$ 와 $s_2(v,z-v)$ 는 각각 임계대역 $z$ 에서의 backward spreading과 forward spreading을 나타내며, 파라미터 $\delta$ 는 MOS 테스트 결과에 최적화 시킬 수 있는 파라미터 이다. 이때 파라미터 $\delta$ 값과 $1+0.02(z-v)dt$ 는 대부분의 psychoacoustic model이 주파수 영역에서의 에너지 확산 현상을 linear convolution process를 통해 수행하고 있으며, 또한 objective evaluation criterion에 최적화 시키기 위해 $\delta\approx 2$ 으로 설정할 수 있고, $1+0.02(z-v)dt\approx 1$ 로 근사화 시킬 수 있다[6][7][8].

그 결과 식(9)는 다음과 같이 간략화 시킬 수 있으며,

$$F_s^e(z;t) \approx \sum_{v=0}^{z-1} s_2(v;z-v)F_s^l(v;t) + \sum_{v=z}^{Z-1} s_1(v-z)F_s^l(v;t) \quad (10)$$

이때 $s_1(v-z)$ 와 $s_2(v,z-v)$ 을 다음과 같은 단일 함수를 이용하여 표현하면

$$ss(v,z) = \begin{cases} s_2(v,b-z), & v < z \\ s_1(v-z), & v \geq z \end{cases} \quad (11)$$

임계대역 내에서의 신호에 대한 excitation에너지는 다음과 같이 표현 할 수가 있다.

$$F_s^e(z;t) = \sum_{v=0}^{Z-1} \left[ ss(v;z)F_s^l(v;t) \right] \quad (12)$$

또한, 이러한 임계대역 내에서의 excitation은 시간 영역에서의 에너지 확산의 영향 또는 post masking spreading을 고려 하여 이전 프레임과 결합되어 지게 되는데, 이 과정을 다음과 같이 표현할 수 있다.

$$F_s^{el}(z;t) = F_s^e(z;t) + T_f(z)F_s^e(z;t-1), \quad \text{where, } 0 \leq z \leq B-1 \quad (13)$$

여기에서 $T_f(z) = e^{-d/\tau(z)}$ 는 신호 에너지의 시간에 따른 확산에 의한 영향을 표현하기 위한 함수이고, $d$ 는 인접 프레임들간의 time distance를 표현하기 위한 파라미터이며, $\tau(z)$ 는 시간영역에서의 마스킹 실험 결과를 수용하기 위한 파라미터 이다[6][7][8].

## 3. Speech Improvement with Noise Perception Pattern Suppression

이와 같이 사람의 청각 시스템에서의 신호 에너지 지각적 응답에 대해 식(7)을 심리 음향적 표현을 통해 다음과 같이 나타낼 수 있다[6].

$$H(z;t) = H\left[ SNR_{post}(z;t) \right] = \left( 1 - \frac{\widehat{F}_d^{el}(z;t)}{F_y^{el}(z;t)} \right) \quad (14)$$

이때, filtering 연산을 간략화 하기 위해 모든 신호 성분의 공통 요소인 attenuation factor $a_0(z)$ 의 영향의 연산을 생략하고, no time-domain memory 상황, 즉, $T_f(z) = 0$ 이라고 가정할 수 있다. 이것은 현재 프레임에 대해서 이러한 filtering을 통해, 잡음 신호 성분의 영향에 대해 지각적으로 투명성을 갖게 된다고 가정을 한다면, noisy frame에 첨가된 잡음 신호 에너지에 의한 시간 영역에서의 프레임간 에너지 확산의 영향을 나타내는 마스킹 현상의 영향을 고려 대상에서 제외 시킬 수 있으며, 음성 신호 에너지에 의한 시간 영역에서의 이러한 영향은 그러한 현상을 포함해서 음성 신호 자체로 지각하게 되므로 음성 신호에 의한 이러한 영향은 고려 대상에서 제외 시킬 수 있다. 실제 유사한 실험 결과에서[6], 이러한 시간 영역에서의 에너지 확산에 의한 영향은 filtering 성능에는 크게 기여하지 않으므로 이러한 가정은 타당성을 갖는다고 할 수 있다.

$$H(z;t) = \left( 1 - \frac{\widehat{F}_d^e(z;t)}{F_y^e(z;t)} \right) \quad (15)$$

이때 *a posteriori SNR*의 산출에 있어, 주파수 차감 형식의 알고리즘의 가장 근본적인 가정인 식(2)로부터, 사람의 청각 시스템이 음성 신호에 대한 잡음에 의한 손상을 지각하게 되는 것은 음성 신호에 지각적인 간섭 효과를 발생 시키는 잡음 신호의 에너지 확산 영향에 의한 것이 된다. 이때 음성 신호 자체의 에너지 확산에 따른 영향은 이러한 신호 자체를 포함하여 신호 자체로 지각하게 되므로, 잡음에 오염된 음성 신호 내에서 잡음 신호 에너지 확산의 영향에 의해 음성 신호가 얼마나 영향을 받았는가를 표현하기 위해 $SNR_{post}(z;t)$ 에 의한 필터, $H^p(z;t)$ 는 다음과 같이 perceptual *a posteriori* SNR에 의해 구성할 수 있다.

$$H^p(z;t) = H\left[ SNR_{post}^p(z;t) \right] = \left( (1+\alpha) - \frac{\widehat{F}_d^e(z;t)}{F_y^l(z;t)} \right) \quad (in\ dB) \quad (16)$$

where,

$SNR_{post}^p(z;t)$: perceptual *a posteri SNR*

이때 $\alpha$ $(0 \leq \alpha \leq 1)$ 값은 에너지 영역( *in dB* )에서 필터의 상대적인 조건 $0 \leq H^p(z;t) \leq 1$ 을 만족시키기 위해 프레임 평균 $SNR_{post}$ 에의해 결정되어지는 scale factor 이다.

또한 잡음의 추정에 있어, 해당 프레임에 첨가된 잡음 신호를 정확히 추정할 수 있다면, 개선된 신호의 경우 원래의 음성 신호와 같은 결과를 얻을 수 있을 것이다. 그러나, 실제로 잡음 신호를 정확하게 추정하는 일은 어려운 일이므로, 추정 오차가 존재하게 되는데, 이러한 추정 오차에 의한 noisy신호 내에서의 영향을 보상해줄 필요가 있다.

이러한 추정한 잡음 신호와 현재 프레임에 포함된 잡음 신호 사이의 잡음 추정 오차의 영향은 묵음 구간의 잡음 신호 프레임을 기준으로 현재 잡음에 오염된 음성 신호 프레임에 대한 unpredictability measure를 통해 추정할 수 있으며, 이러한 잡음 추정 오차에 대한 noisy 프레임 내 음성 신호 성분에 대한 영향은 다음과 같이 표

할 수 있다.

$$F_y^{cl}(z;t) = c_N(z;t) \cdot F_y^{l}(z;t) \qquad (17)$$

여기에서 $c_N(z;t)$ 는 noise unpredictability measure를 나타내며, 이러한 noisy 신호 내에서의 잡음 추정 오차에 의한 영향은 다음과 같이 추정 오차에 의한 $SNR_{post}$ 과 같이 표현할 수 있다.

$$G(z;t) = \frac{F_y^{ce}(z;t)}{F_y^{e}(z;t)} \quad \text{where, } 0 \le G(z;t) \le 1 \qquad (18)$$

이와 같이 잡음 신호에 대한 사람의 청각 시스템에서의 지각 패턴 제어를 통한 음성 개선 시스템은 다음과 같이 구성 할 수가 있다.

$$\widehat{F}_s^{l}(z;t) = G(z;t) \cdot H^{p}(z;t) \cdot F_y^{l}(z;t) \qquad (19)$$

## 4. 실험 및 결론

제안된 음성 개선 알고리즘의 성능 평가를 위해 AWGN 잡음 환경에서 10 dB SNR을 갖도록 잡음에 오염 시킨 음성 신호를 이용하여 기타 심리음향적 특성을 이용한 음성 개선 알고리즘들과 비교를 하였다.

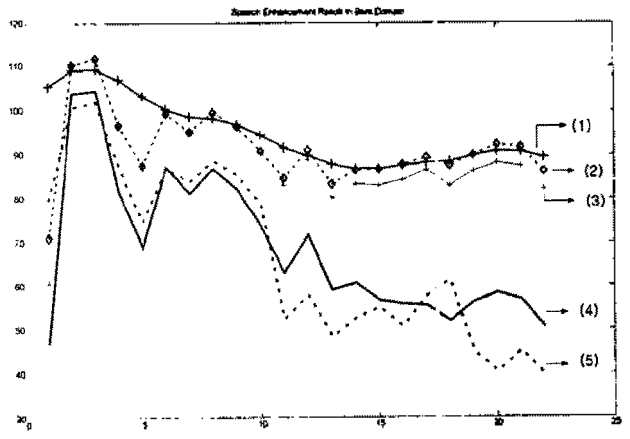


그림 1. Speech Enhancement Results in Bark Domain

그림 1.은 임의의 한 프레임 구간에 대한 Bark domain 에서의 음성 개선 알고리즘의 처리 결과에 에너지 스펙트럼을 나타낸다. 그림에서 (1)의 결과는 psychoacoustic noise shaping 방법[6]을, (2)는 noisy 신호를, (3)은 masking property에 근거해서 차감 파라미터를 적응 시킨 방법[5]을, (4)는 제안한 방법을, (5)는 clean speech signal의 에너지 스펙트럼의 결과를 나타낸다. 그림에서 확인 할 수 있듯이 제안한 알고리즘의 경우 음성 개선 알고리즘의 적용 후 고주파 영역에서 여전히 과도하게 존재하게 되는 잡음 에너지 제어에 효과적이며, 각 주파수 대역 내에서 원 음성 신호와 유사한 에너지 분포를 갖는 결과를 나타내었으며 SNR과 MOS 실험 결과에서도 높은 향상 능력을 나타내었다.

본 논문에서는 사람의 청각 시스템에서의 지각 패턴을 이용한 주파수 차감 기법 기반의 필터링을 통한 음질 향상 알고리즘에 대한 연구를 수행 하였다. 이것은 오염된 음성 신호 내의 추가된 잡음 에너지의 추정을 통해 잡음 에너지 확산에 의한 음성 신호 에너지에 의한 지각적인 에너지 간섭 현상을 제어(suppression)하고 잡음 에너지 추정 오차에 의한 손상된 음성 신호내의 순음(tonal) 특성과 비순음(non-tonal) 특성을 noise unpredictability를 통해 보상 시켜 주는 방식이다.

## 참고문헌


1. S. F. Boll, "Suppression of acoustic noise speech using spectral subtraction"' IEEE Trans. Acoustic., Speech, Signal Processing, ASSP-27, 113-120, Apr. 1979.
2. P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars," Speech Commun., 11, 215-228, June 1992.
3. R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoustic., Speech, Signal Processing, 28, 137-145, Apr. 1980.
4. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoustic., Speech, Signal Processing, 32, 1109-1121, Dec. 1984.
5. Nathalie Virag "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," IEEE Trans. Acoustic., Speech, Signal Processing, Vol. 7 N.2, March, 1999.
6. Dionysis E. Tsoukalas, John Mourjopoulos, George Kokkinakis, "Perceptual Filter for Audio Signal Enhancement", J. Audio Eng. Soc. Vol. 45 No.1/2, January, 1997.
7. E. Zwicker, H. Fastl, *Psychoacoustics : Facts and Models*, Springer 2nd Edition, 1999.
8. C, J. Moore, *Hearing*, Academic Press, 1995
9. John R. Deller, Jr., John G. Proakis, John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Cimpany, 1993