한국음향학회지 제20권 제3호 pp. 48~55 (2001)

시간-주파수 영역에서 음성/잡음 우세 결정에 의한 새로운 잡음처리

A Novel Speech Enhancement Based on Speech/Noise-dominant Decision in Time-frequency Domain

윤 석 현*, 유 창 동*
(Sukhyun Yoon*, Chang D. Yoo*)

*한국과학기술원 전기 및 전자공학과

(접수일자: 2000년 12월 13일; 수정일자: 2001년 2월 5일; 채택일자: 2001년 2월 27일)

가산적이고 비정상적인 잡음을 줄이는 새로운 방법이 제안되었다. 본 방법은 잡음에 대한 정보나 묵음구간에서의 잡음추정을 필요로 하지 않는다. 잡음처리는 각 시간 프레임에서 주파수대역을 기본으로 하여 수행된다. 어떤 프레임에서 특정한 주파수대역이 음성이 우세한지 혹은 잡음이 우세한지에 대한 결정과 인간청각기의 매스킹 성질을 기반으로 하여, 적절한 양의 잡음을 주파수 차감법을 이용하여 제거한다. 제안된 방법은 다양한 환경에서 (자동차 잡음, F16 잡음, 백색 잡음, 핑크 잡음, 탱크 잡음, 혼선잡음) 성능평가가 이루어졌다. 그리고 일반적인 주파수차감법과 비교하여 세그멘탈 신호대 잡음비 (SNR)를 구하고, 시각적 측정 척도인 스펙트로그램과 듣기 평가를 통해, 음성왜곡은 줄이면서 효과적으로 잡음을 줄일 수 있음을 알 수 있다.

핵심용어: 잡음처리, 음성/잡음-우세결정, 분류, 매스킹
투고분야: 음성처리 분야 (2,3)

A novel method to reduce additive non-stationary noise is proposed. The method requires neither the information about noise nor the estimate of the noise statistics from any pause regions. The enhancement is performed on a band-by-band basis for each time frame. Based on both the decision on whether a particular band in a frame is speech or noise dominant and the masking property of the human auditory system, an appropriate amount of noise is reduced using spectral subtraction. The proposed method was tested on various noisy conditions (car noise, F16 noise, white Gaussian noise, pink noise, tank noise and babble noise) and on the basis of comparing segmental SNR with spectral subtraction method and visually inspecting the enhanced spectrograms and listening to the enhanced speech, the method was able to effectively reduce various noise while minimizing distortion to speech.

Keywords: *Speech enhancement, Speech/noise-dominant decision, Classification, Masking*
Ask subject classification: *Speech signal processing (2,3)*

I. 서 론

다양한 음성처리 시스템의 꾸준한 성장으로 단채널에서의 고성능 잡음처리 시스템의 필요성은 나날이 증대되고 있다. 음성에 잡음이 부가되면 음성인식기의 인식률이 떨어지고 음성부호기 (vocoder)의 코딩 효율성을 낮춘다. 일반적으로, 잡음은 명료성을 떨어뜨리고 듣는 사람에게 귀의 피로감을 안겨준다. 이러한 점을 직시하여 잡음을 줄이기 위한 다양한 방법이 제안되어 왔다: 주파수차감법에

책임저자: 윤석현 (ronald@eeinfo.kaist.ac.kr)
305-701 대전시 유성구 구성동 373-1
한국과학기술원 전기 및 전자공학과 Multimedia Processing Lab.
(전화: 042-869-5470; 팩스: 042-862-0559)

기반한 방법[1-4], 소프트 디시즌 (soft-decision) 필터링 방법[5], 최소평균제곱오차 추정 (MMSE estimation) 방법[6,7], 음성모델에 기반한 음성잡음처리방법[8-10], 그리고 인간 청각기 특성을 이용한 잡음처리방법[11-14]. 위의 방법들은 잡음을 제거하기 위해서 잡음에 대한 통계적인 정보가 필요하고, 만일 그 정보를 알 수 없을 때는 보통 묵음구간을 통해 잡음을 추정한다. 하지만 정확한 묵음구간 검출은 매우 어려울 뿐더러 잘못된 묵음구간 검출은 오히려 음성을 더욱 왜곡시키는 결과를 낳는다. 본 방법은 잡음에 대한 어떠한 정보나 묵음구간 검출을 필요로 하지 않는다. 과거에도 묵음구간 검출 없이 잡음을 줄이려는 시도가 있었으나[15,16] 그 시도들은 만족스러운 결과를 내놓지 못했다.

본 논문에서는 하나의 시간 프레임에서 각각의 주파수 대역별로 잡음처리를 수행한다. 어떤 프레임의 특정한 대역에 속하는 주파수 성분 크기의 합과 그 대역에 해당하는 과거의 주파수 성분 크기의 합들과 비교하여, 그 프레임의 특정한 대역에서 음성성분이 대부분을 차지하는지 (음성우세) 잡음성분이 대부분을 차지하는지 (잡음우세)에 대한 결정을 하고 인간 청각기의 매스킹 성질에 기반하여, 음성왜곡은 최소로 하고 잡음을 최대로 제거한다. 좀더 구체적으로 말하자면, 과거 프레임들의 특정 대역에 해당하는 합들을 오름차순으로 정렬하여, 그 모양에 따라 대역을 두 가지 종류로 분류한다. 그리고 종류에 따라 다른 기준을 적용하여 오름차순으로 정렬된 합들과 비교함으로써, 현재 처리하려는 프레임의 각각의 대역에서 음성성분이 대부분을 차지하는지 (음성우세) 잡음성분이 대부분을 차지하는지 (잡음우세) 결정을 한다. 이러한 결정과 인간 청각기의 매스킹 성질에 기반하여, 각각의 음성/잡음우세 대역에서 알맞은 양의 잡음을 주파수 차감법을 이용하여 제거한다.

본 논문은 다음과 같이 이루어졌다. 제2장에서는 임계 밴드 (critical band)와 매스킹 성질에 관해 설명을 한다. 제3장에서는 전체 잡음처리 시스템을 설명한다: 3.1절에서는 어떻게 각 대역에 따라 정렬된 수열을 얻을 수 있는지 설명한다. 3.2절에서는 각각의 정렬된 수열을 어떻게 두 종류로 분류할 수 있는지 나타낸다. 3.3절에서는 어떻게 잡음 스펙트럼을 추정하는지 설명한다. 제4장에서는 제안된 방법을 이용한 실험적인 결과와 몇몇의 예를 보여준다. 마지막으로 제5장에서는 내용을 재정리하고 결론을 내린다.

II. 임계 밴드 (critical band)와 매스킹 성질

2.1. 임계 밴드 (critical band)

내이(內耳)에서 소리의 주파수 성분과 달팽이관의 특정 위치와 서로 매칭 (frequency-to-place transformation)이 되면서 소리를 사람이 인지할 수 있게 된다[19]. 어떤 하나의 주파수 성분이 주어지면, 귀의 달팽이관의 어떤 특정 부분이 활성화될 것이다. 이 때, 임계 밴드 (critical band)는 달팽이관의 동일한 부분을 활성화시키는 그 주어진 주파수 성분 주위의 주파수 성분들을 포함하는 최소의 밴드를 뜻한다. 일정한 음압 수준 (sound pressure level)을 유지하는 협대역 잡음원이 있다고 하자. 이 잡음원의 대역폭을 넓히더라도 임계 대역폭 (critical bandwidth) 내에 있는 동안은 그 소리를 느끼는 강도 (perceived loudness)가 일정하게 유지된다. 하지만 그 잡음원의 대역폭이 임계 대역폭을 넘어서는 순간, 느끼는 강도 (perceived loudness)는 증가하게 된다[19]. 이런 실험을 통해 임계 밴드를 정할 수 있다. 표 1은 임계 밴드의 번호와 중심주파수 그리고 대역폭을 나타내었다[20]. 본 논문은

표 1. 임계 밴드 (critical band)의 번호와 중심주파수 그리고 대역폭을 나타내었다
Table 1. The band number, center frequency and bandwidth in critical band.

Band No.	Center Freq.(Hz)	Bandwidth(Hz)	Band No.	Center Freq.(Hz)	Bandwidth(Hz)	Band No.	Center Freq.(Hz)	Bandwidth(Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

각 프레임의 임계 밴드를 중심으로 하여 음성을 처리한다.

2.2. 매스킹 성질

매스킹은 하나의 소리가 다른 소리의 존재로 인해 들리지 않게 되는 과정이다[21]. 본 논문에서는 주파수 (simultaneous) 매스킹을 사용하였다. 주파수 매스킹은 두 개 혹은 그 이상의 주파수 성분이 동시에 나타났을 때, 어떤 특정 주파수 성분이 있음에도 불구하고 이웃한 주파수 성분의 영향을 받아 귀에는 그 주파수 성분이 들리지 않게 되는 과정이다[21]. 잡음처리의 문제점 중의 하나는 잡음을 없애는 과정에서 음성성분까지 제거가 되어 명료성이 떨어지는 현상이다. 따라서 최근에는 매스킹되어 들리지 않을 부분은 잡음추정치보다 적게 빼주는 방식으로 최대한 음성성분을 보존하려는 노력을 하고 있다[11-14]. 본 논문에서도 음성성분이 대부분이라 생각되는 부분(음성우세지역)은 작은 양의 잡음추정치를 빼줌으로써 음성성분을 최대한 살리는 방법을 채택하고 있다.

III. 잡음처리 시스템

본 방법은 다음 3가지 단계를 거친다. 첫째, 잡음이 섞인 음성 $y[n]$을 $w[n]$으로 창을 씌운다. 창으로 씌워진 신호를 DFT (Discrete Fourier Transform) 계수로 변환한다. 각 임계 밴드 안에 있는 계수의 크기들을 합한다. 그리고 과거 L개의 프레임 동안 각 임계 밴드에서의 합들을 오름차순으로 정렬한다. 둘째, 근사함수를 이용하여 정렬된 수열을 모양에 따라 두 종류로 분류한다. 종류에 따라 다른 기준을 적용하여 오름차순으로 정렬된 합들과 비교함으로써, 현재 처리하려는 프레임의 각각의 대역에서 음성성분이 대부분을 차지하는지 (음성우세) 잡음성분이 대부분을 차지하는지 (잡음우세) 결정을 한다. 이러한 결정과 인간청각기의 매스킹 성질에 기반하여, 각각의 음성/잡음우세대역에서 알맞은 양의 잡음을 주파수 차감법을 이용하여 제거한다. 그림 1은 전체 시스템을 보여준다.

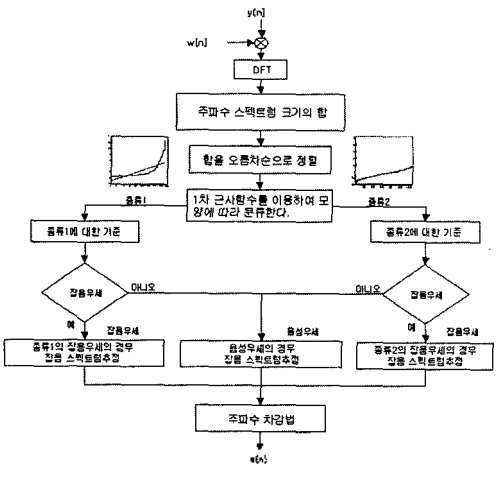

그림 1. 전체시스템
Fig. 1. The overall system.

3.1. 오름차순 정렬

n번째 프레임의 i번째 임계 밴드 안에 있는 주파수 크기를 합한다. 그 합 $A[i,n]$은 다음과 같다.

$$A[i,n]=\sum_{k\in CB_i}|Y[k,n]| \tag{1}$$

단, CB_i는 i번째 임계 밴드에 속하는 주파수 성분의 집합이다. 그리고 $Y[k,n]$은 잡음이 섞인 음성의 n번째 프레임의 k번째 DFT (Discrete Fourier Transform) 계수를 나타낸다. 모든 i마다 길이 L인 수열 $\{A[i,j]\}_{j=n-L}^{n-1}$을 오름차순으로 정렬하여 수열 $\{E[i,j]\}_{j=1}^{L}$를 얻을 수 있다. $E[i,q]$은 수열 $\{A[i,j]\}_{j=n-L}^{n-1}$의 q번째로 큰 항이다. 예를 들어, $E[i,1]=\min_j(A[i,j])$, $E[i,L]=\max_j(A[i,j])$ for $n-L\le j\le n-1$. L값이 작으면 작을수록 변화하는 잡음에 더 빨리 적응할 수 있다. 하지만, L값이 너무 작아지면 잡음 추정이 부정확해진다. 그림 2는 $i=5$, 10, 15에 대해서 정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$의 그래프이다.

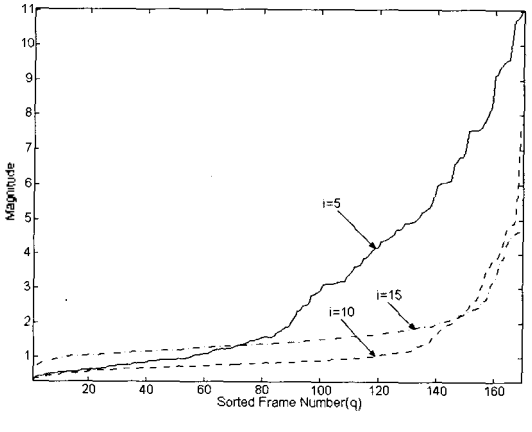

그림 2. $i=5$, 10, 15에 대해서 정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$의 그래프이다 위의 그림에서 L=170이다
Fig. 2. The plot of ordered sequences $\{E[i,j]\}_{j=1}^{L}$ for $i=5$, 10, 15 where L=170.

3.2. 정렬된 수열의 분류

정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$은 모양에 따라 두 종류로 분류한다. 최근 L프레임 동안 상당한 양의 음성이 있었다면 $\{E[i,j]\}_{j=1}^{L}$의 그래프는 휘어진 형태이다. 이는 잡음성분과 음성성분의 명확한 차이를 보여준다. 이 경우에 $\{E[i,j]\}_{j=1}^{L}$는 '종류1'로 분류된다. 최근 L 프레임 동안 오직 잡음만 있거나 작은 크기의 음성성분만 있었다면, $\{E[i,j]\}_{j=1}^{L}$의 그래프는 수평한 형태를 하고 있을 것이다. 이는 잡음과 음성의 구분이 어려움을 나타낸다. 이 경우에 $\{E[i,j]\}_{j=1}^{L}$는 '종류2'로 분류된다.

위의 그래프를 1차함수로 근사화한다[17]. 그리고 1차 함수의 y절편을 이용하여 정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$을 '종류1' 혹은 '종류2'로 분류한다. 만일 y절편이 0보다 작다면, $\{E[i,j]\}_{j=1}^{L}$는 '종류1'로 분류된다. 만일 y절편이 0보다 크다면, $\{E[i,j]\}_{j=1}^{L}$는 '종류2'로 분류된다. 그림 3은 $i=5,15$일 때 정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$의 '종류1'과 '종류2'로의 분류과정을 보여준다.

3.3. 잡음 스펙트럼의 추정

최근 L개의 프레임에서 오름차순으로 정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$을 $A[i,n]$과 비교함으로써 $(i,n)-region$[1]

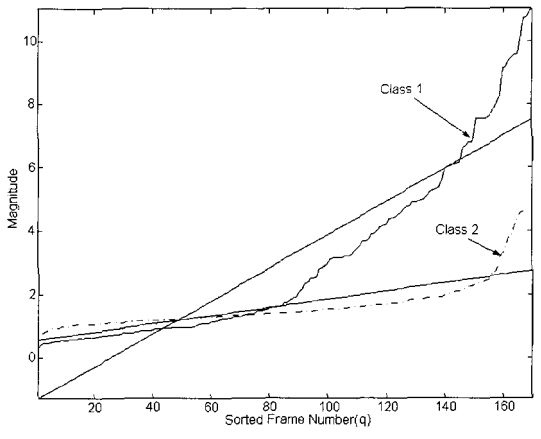

그림 3. $i=5,15$일 때 1차함수를 이용하여 정렬된 수열 $\{E[i,j]\}_{j=1}^{L}$의 '종류1'과 '종류2'로의 분류

Fig. 3. The classification of the ordered sequences $\{E[i,j]\}_{j=1}^{L}$ for $i=5,15$ where L=170 into either Class 1 or Class 2 by fitting a first order polynomial.

1) $(i,n)-region$는 n번째 프레임의 i번째 임계 밴드(critical band)를 의미한다.

2) 표시 $\lceil X \rceil$ 는 양의 무한대방향으로 X와 가장 가까운 정수를 나타낸다. 예를 들어, $\lceil 1.2 \rceil = 2$

이 음성우세인지 잡음우세인지 결정한다. 만일 $(i,n)-region$이 음성우세로 결정이 되면, 음성성분을 보존하기 위해 $\{E[i,j]\}_{j=1}^{L}$내에서 작은 값을 사용하여 $|Y[k,n]|$ (단, $k\in CB_i$)으로부터 빼준다. 이 경우에 남아있는 잡음성분은 음성성분에 의해 매스킹되어 큰 영향을 끼치지 못한다. 만일 $(i,n)-region$이 잡음 우세로 결정되면, $\{E[i,j]\}_{j=1}^{L}$내에서 큰 값을 사용하여 $|Y[k,n]|$ (단, $k\in CB_i$)으로부터 빼준다. 잡음 스펙트럼 $|M[i,n]|$는 각 종류에 따라 다른 기준을 적용함으로써 얻을 수 있다. 우리는 $|Y[k,n]|$으로부터 $|M[i,n]|$을 빼줌으로써 깨끗한 음성을 얻을 수 있다. 수학적인 표현은 다음과 같다.

$$S[k,n]=rect(|Y[k,n]|-|M[i,n]|) \quad k\in CB_i \quad (2)$$

단, $S[k,n]$은 잡음처리된 음성의 스펙트럼성분의 크기를 나타내고 $rect(\cdot)$는 반파정류기를 나타낸다.

3.3.1. '종류1'에 대한 기준

$\{E[i,j]\}_{j=1}^{L}$가 '종류1'로 분류가 된다면 최근 L프레임 동안 i번째 임계 밴드내에는 강한 음성성분이 있음을 나타낸다. '종류1'로 분류된 대역이 고주파에서의 (예를 들어, $i>17$) 대역이라면 저주파에서 (예를 들어, $i\le 17$) '종류1'로 분류된 대역보다는 강한 음성성분이 그리 많지는 않다. 따라서 '종류1'로 분류된 경우에는 저주파에서보다는 고주파에서 음성과 잡음의 차이가 더 명확하게 보인다. 저주파에는 오히려 강한 음성성분이 많아서 잡음을 분별하기가 어렵다. 만일 고주파에서 $A[i,n]$이 $\{E[i,j]\}_{j=1}^{L}$의 평균보다 작다면, 즉, $A[i,n]<\frac{1}{L}\sum_{q=1}^{L}E[i,q]$이면, 그때 $(i,n)-region$은 잡음우세이다. 이 $region$에서 $A[i,n]$이 작은 값이므로 음성성분이 있다 할지라도 큰 영향을 미치는 성분은 아닐 것이다.

저주파에서 $(i,n)-region$이 묵음구간에 있다면 $A[i,n]$은 상대적으로 작은 값을 갖게 된다. 따라서 만일 $A[i,n]$을 $\{E[i,j]\}_{j=1}^{L}$내의 값과 비교하여 $A[i,n]$이 상대적으로 작다면, 다시 말해, $E[i,\lceil L\cdot a\rceil]$[2] 보다(단, a는 0.25에서 0.35사이의 값이다.) 작다면, 그때 $(i,n)-region$은 잡음우세이다.

위에 언급된 경우에는 잡음 스펙트럼 $|M[i,n]|$은 $\{E[i,j]\}_{j=1}^{L}$내에서 큰 값으로 추정된다. 그렇지 않으면, $|M[i,n]|$은 $\{E[i,j]\}_{j=1}^{L}$내에서 작은 값으로 추정된다. 잡음 스펙트럼은 다음과 같이 추정된다.

만일 (고주파에서 $A[i,n] < \frac{1}{L}\sum_{q=1}^{L} E[i,q]$) 혹은
(저주파에서 $A[i,n] < E[i, \lceil L \cdot a \rceil]$)이면, 그때

$(i,n)-region$ 는 잡음우세

$\Rightarrow |M[i,n]| = E[i, \lceil L \cdot high \rceil]/B_i \quad high \in [0.9, 1]$

그렇지 않으면,

$(i,n)-region$ 는 음성우세

$\Rightarrow |M[i,n]| = E[i, \lceil L \cdot low \rceil]/B_i \quad low \in [0.25, 0.35]$

단, B_i는 CB_i안에 있는 주파수성분 (frequency bin)의 개수이다. 그리고 a는 0.25에서 0.35의 값을 가진다.

3.3.2. '종류2'에 대한 기준

$\{E[i,j]\}_{j=1}^{L}$이 '종류2'에 속한다면 이 때는 최근 L프레임동안 i번째 임계 밴드는 잡음이나 약한 음성성분이 대부분이다. 하지만, 만일 $A[i,n]$이 $\{E[i,j]\}_{j=1}^{L}$내의 값과 비교하여 특별히 큰 값이라면, $(i,n)-region$은 음성우세라고 볼 수 있다. 따라서

만일 ($A[i,n]$이 고주파에 있거나) 혹은
(저주파에서 $A[i,n] < E[i, \lceil L \cdot b \rceil]$)이면, 그때

$(i,n)-region$ 는 잡음우세

$\Rightarrow |M[i,n]| = c \cdot E[i,L]/B_i \quad c \in (1,2]$

그렇지 않으면,

$(i,n)-region$ 는 음성우세

$\Rightarrow |M[i,n]| = E[i, \lceil L \cdot low \rceil]/B_i \quad low \in [0.25, 0.35]$

단, b의 범위는 0.9에서 1이다.

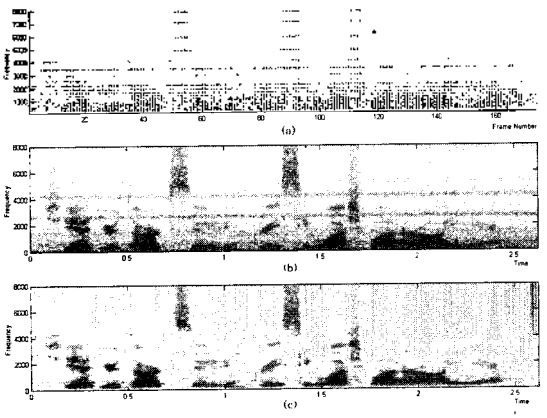

그림 4. (a)는 제안된 방법에 의해 결정된 시간-주파수 영역에서의 음성우세지역을 나타내었다. 그림 4(b)와 (c)는 각각 F16잡음에 의해 오염된 음성(SNR=10dB)과 잡음처리된 음성의 스펙트로그램

Fig. 4. (a)The plot shows speech dominant part in time-frequency domain as determined by the method. The spectrograms of (b)noisy speech degraded by F16 noise(SNR=10dB) and (c)enhanced speech.

그림 4에서 (a)는 위의 방법에 의해 결정된 시간-주파수 영역에서의 음성우세지역을 나타내었다. 그리고 그림 4(b)와 (c)는 각각 F16잡음에 의해 오염된 음성 (SNR=10dB)과 잡음처리된 음성의 스펙트로그램을 보여준다.

IV. 평 가

제안된 방법의 성능을 보여주기 위해 우리는 다양한 잡음상황에서 평가를 해보았다. 음성문장은 TIMIT 데이터베이스에서 선택되었으며 문장은 다음 두 문장이다. - "She had your dark suit in greasy wash water all year" 와 "Scholastic aptitude is judged by standardized tests" - 각각 남성과 여성에 의해 읽혀진 문장이다. 매개변수는 다음 값으로 선택되었다: 1)50% 오버랩을 하면서 길이 N=512인 해밍 창 (Hamming window); 2)임계 밴드의 개수는 22개; 3)a=0.3, b=0.9, c=2, high=0.9, low=0.3; 4)그림 5(c)와 그림 6(c)는 L=100, 그림 7(b)는 L=50이 사용되었다.

Noisex-92 데이터베이스에서 6개의 다른 잡음을 뽑아내어 평가하는데 사용하였다. 6개의 잡음은 자동차잡음, F16잡음, 백색잡음, 핑크잡음, 탱크잡음, 혼선잡음이다.

스펙트로그램은 시각적 측정 도구로서 가로축은 시간,

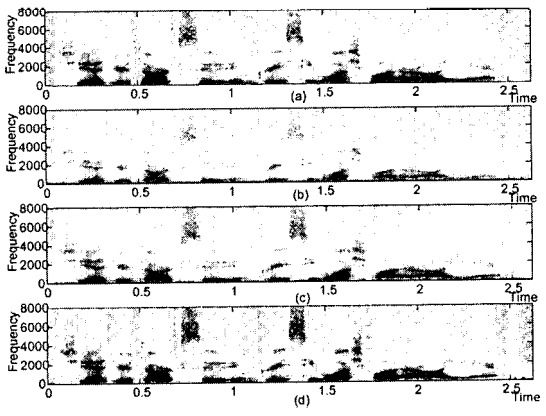

그림 5. (a)깨끗한 음성 ("She had your dark suit in greasy wash water all year" 남자의 음성) (b)자동차잡음에 의해 오염된 음성(SNR=-5dB) (c)묵음구간에서의 잡음추정을 이용한 주파수차감법에 의해 잡음처리된 음성 그리고 (d)제안된 방법에 의해 잡음처리된 음성의 스펙트로그램

Fig. 5. The spectrograms of (a)clean speech ("She had your dark suit in greasy wash water all year" spoken by a male speaker), (b)noisy speech degraded by car noise(SNR=-5dB), (c)enhanced speech by spectral subtraction with pause detection and (d)enhanced speech by the proposed method.

세로축은 주파수를 나타낸다. 시각적으로 음성의 상태를 쉽게 알 수 있다는 장점이 있다.

그림 5(a), (b), (c), (d)는 각각 깨끗한 음성 ("She had your dark suit in greasy wash water all year" 남자의 음성), 자동차잡음에 의해 오염된 음성 (SNR=-5dB), 묵음구간에서의 잡음추정을 이용한 주파수차감법에 의해 잡음처리된 음성, 그리고 제안된 방법에 의해 잡음처리된 음성의 스펙트로그램을 나타낸다. 그림 5(b)를 보면 자동차잡음의 특성은 저주파 (100Hz부근)에 대부분의 잡음성분이 몰려 있는 특성이 있다. 그 외의 부분은 백색잡음처럼 넓게 고르게 퍼져 있다. 그림 5(c)는 묵음구간을 이용한 잡음처리를 통해 나오는 스펙트로그램인데, 깨끗한 음성의 스펙트로그램을 나타내는 그림 5(a)와 비교해서 그리 선명한 모습을 보여 주지 못 한다.(저주파 부근을 자세히 보면 잡음이 잘 지워지지 않았음을 볼 수 있다.) 그림 5(d)를 보면 그림 5(c)에 비해 선명하고 저주파 부근에 잡음이 거의 사라졌음을 관찰할수 있다. 그림 6은 각각 (a)깨끗한 음성 ("Scholastic aptitude is judged by standardized tests" 여자의 음성), (b)탱크잡음에 의해 오염된 음성 (SNR=10dB), (c)묵음구간에서의 잡음추정을 이용한 주파수차감법에 의해 잡음처리된 음성 그리고 (d)제안된 방법에 의해 잡음처리된 음성의 스펙트로그램을 나타낸다. 그림 6(b)를 보면 탱크잡음은 저주파 (1kHz 아래부분)부근에 강하게 나타나고 그 외의 부분은 저주파보다는 약한 형태로 나타난다. 그림 6(c)를 보면 저주파부근에 잡음성분은 어느 정도 제거하였으나 그 외의 부분은 제대로 처리하지 못 하였음을 볼 수 있다. 이는 묵음구간만을 통해 잡음을 추정하므로 변화하는 잡음에 빠르게 대처하지 못 하기 때문이다. 제안된 방법에 의해 나온 그림 6(d)를 보면 탱크잡음의 저주파성분 (1kHz 아래부분)뿐만 아니라 그외

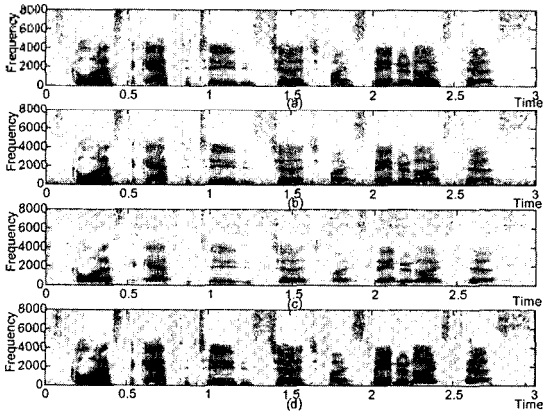

그림 6. (a)깨끗한 음성 ("Scholastic aptitude is judged by standardized tests" 여자의 음성) (b)탱크잡음에 의해 오염된 음성(SNR=10dB) (c)묵음구간에서의 잡음 추정을 이용한 주파수차감법에 의해 잡음처리된 음성 그리고 (d)제안된 방법에 의해 잡음처리된 음성의 스펙트로그램

Fig. 6. The spectrograms of (a)clean speech ("Scholastic aptitude is judged by standardized tests" spoken by a female speaker), (b)noisy speech degraded by tank noise (SNR=10dB), (c)enhanced speech by spectral subtraction with pause detection and (d)enhanced speech by the proposed method.

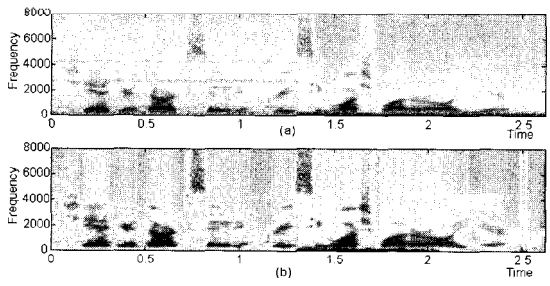

그림 7. (a)과 (b)는 각각 잡음에 의해 오염된 음성과 잡음처리된 음성의 스펙트로그램을 보여준다. 이때는 처음 1.3초동안 F16잡음(SNR=10dB)으로 오염을 시키고 다음 1.3초 동안은 car잡음(SNR=-5dB)으로 오염을 시켰다

Fig. 7. The spectrograms of (a)noisy speech and (b)enhanced speech. Noisy speech was obtained by degrading clean speech with F16 noise (SNR=10dB) for the duration 1.3 seconds followed by car noise (SNR=-5dB) for another 1.3 seconds.

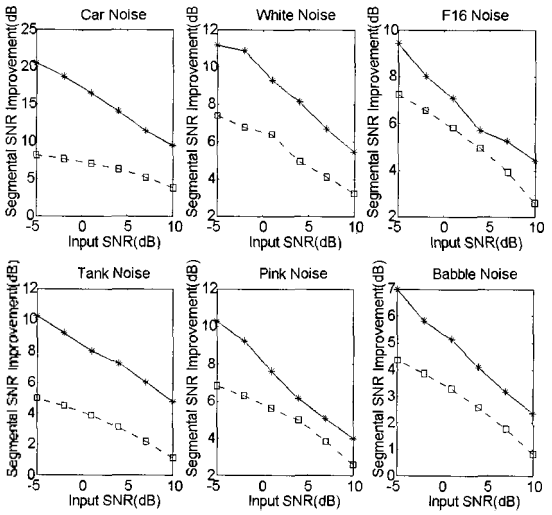

그림 8. 입력SNR대 향상된 segmental SNR의 그래프를 나타내었다 - 자동차 잡음, F16 잡음, 백색 잡음, 핑크 잡음, 탱크 잡음, 혼선잡음- 제안된 방법과 묵음구간추정을 이용한 주파수차감법을 비교하여 나타내었다 실선 : 제안된 방법, 점선 : 묵음구간에서의 잡음추정을 이용한 주파수차감법

Fig. 8. The plots of segmental SNR improvement versus initial SNR for various noise - car noise, white noise, F16 noise, tank noise, pink noise and babble noise. solid line : proposed method, dashed line : spectral subtraction with noise estimation using pause detection.

표 2. 듣기평가결과
Table 2. Result of listening.

방법	F16잡음 (5dB)	F16잡음 (10dB)	탱크잡음 (5dB)	탱크잡음 (10dB)
묵음구간을 이용한 주파수차감법	2.67	2.93	2.53	2.8
제안된 방법	3.1	3.7	3.3	3.8

의 구역도 매우 깨끗하게 잡음이 제거되어 있고 음성성분도 잘 보존되어 있음을 알 수 있다. 그림 7(a)와 (b)는 각각 잡음에 의해 오염된 음성과 제안된 방법에 의해 잡음처리된 음성의 스펙트로그램을 보여준다. 이때는 처음 1.3초동안 F16잡음 (SNR=10dB)으로 오염을 시키고 다음 1.3초동안은 car잡음 (SNR=-5dB)으로 오염을 시켰다. 그림 7은 갑작스럽게 잡음의 특성이 변하는 상황에서도 빠르게 적응해 나가는 모습을 보여준다. 기존의 방법에서는 잡음의 특성이 매우 느리게 변화하거나 혹은 정상적 (stationary)이라는 가정을 하는 것이 대부분이다[1-14]. 하지만 본 방법은 갑작스럽게 잡음특성이 변하는 상황에서도 빠르게 적응하여 잡음제거를 한다. 그림 8은 입력SNR대 향상된 세그멘탈 (segmental) SNR의 그래프를 나타내었다 - 자동차 잡음, F16 잡음, 백색 잡음, 핑크 잡음, 탱크 잡음, 혼선잡음에 대해서 제안된 방법과 묵음구간추정을 이용한 주파수차감법을 각각 비교하여 나타내었다. 세그멘탈 SNR은 객관적 (objective) 평가방법으로서 음성의 질이 어느 정도 향상되었는지에 대한 척도가 된다. 그림 8을 보면 다양한 잡음상황에서도 기존의 방법보다 전체적으로 향상된 성능을 보인다. 특히, 자동차잡음에 대해서는 기존의 방법에 비해 탁월한 성능을 보이고 있음을 알 수 있다.

세그멘탈 SNR이 좋다고 듣기평가에서 항상 좋은 결과를 낸다고 볼 수 없다[22]. 따라서 본 논문에서는 듣기평가도 하였다. 듣기 평가는 10명의 사람 (listeners)이 참여하였다. F16잡음과 탱크잡음에 의해 오염된 음성을 시험해 보았다. 각 사람은 1점에서부터 5점까지의 점수로 음질을 평가하게 된다. - 음질이 좋을수록 높은 점수를 준다. 음질평가에는 배경잡음과 음성왜곡을 고려한다. 테스트는 DAT-tape에 녹음된 음성을 헤드폰으로 들으면서 이루어진다. 각 사람에게 먼저 깨끗한 음성과 잡음이 섞인 음성을 2번 반복하여 먼저 들려준 다음, 묵음구간에서의 잡음추정을 이용한 주파수차감법에 의해 잡음처리된 음성과 제안된 방법에 의해 잡음처리된 음성을 3번 반복하여 들려준다. 이 때 사람들에 의해 점수가 매겨지고 듣기가 끝난 후에는 평균을 내어 최종점수를 낸다. 표 2는 듣기평가의 결과를 보여준다. 듣기평가에서도 제안된 방법은 좀더 향상된 결과를 보여준다.

V. 결 론

본 논문에서 우리는 잡음에 대한 사전정보나 묵음구간에서의 잡음추정을 필요로 하지 않는 새로운 잡음처리기법을 소개하였다. 잡음처리는 각 시간 프레임에서 주파수대역을 기본으로 하여 수행된다. 어떤 프레임에서 특정한 주파수대역이 음성이 우세한지 혹은 잡음이 우세한지에 대한 결정과 인간청각기의 매스킹 성질을 기반으로 하여, 각각의 음성/잡음우세대역에서 알맞은 양의 잡음을 주파수 차감법을 이용하여 제거한다.

제안된 방법은 다양한 환경에서 성능평가가 이루어졌다.(자동차 잡음, F16 잡음, 백색 잡음, 핑크 잡음, 탱크 잡음, 혼선잡음) 그리고 일반적인 주파수차감법과 비교하여 세그멘탈 신호대 잡음비를 구하고, 시각적 측정 척도인 스펙트로그램과 듣기 평가를 통해, 음성왜곡은 줄이면서 효과적으로 잡음을 줄일 수 있음을 알 수 있다.

참 고 문 헌

1. S. F. Boll, "Suppression of acoustic noise speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.
2. J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE* vol. 67, pp. 1586-1604, Dec. 1979.
3. M. erouti and R. chwartz, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE ICASPP*, Washington, DC, pp. 208-211, Apr. 1979.
4. P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor(NSS), hidden Markov models and projection, for robust recognition in cars," *Speech Commun.*, vol. 11, pp. 215-228, June, 1992.
5. R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 137-145, Apr., 1980.
6. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109-1121, Dec., 1984.
7. Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 443-445, Apr., 1985.

8. Y. Ephraim, "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models," *IEEE Trans on Signal Processing*, vol. **40**, pp. 725-735, Apr. 1992.
9. Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *IEEE proceeding*, vol. **80**, pp. 1526-1555, Oct., 1992.
10. H. Sameti, H. Sheikhzadeh and L. Deng, "HMMBased Strategies for Enhancement of Speech Signals Embedded in Nonstationary noise," *IEEE tran. on speech and audio processing*, vol. **6**, pp. 445-455, Sep., 1998.
11. D. Tsoukalas, M. Paraskevas and J. Mourjopoulos, "Speech enhancement using psycho-acoustic criteria," proc. IEEE ICASSP, pp. 359-361, Apr., 1993.
12. T. Usagawa and M. Iwata and M. Ebata, "Speech parameter extraction in noisy environment using a masking model," proc. IEEE ICASSP, pp. 81-84, Apr., 1994.
13. S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constrains on an auditory-based spectrum," *IEEE tran. on speech and audio processing*, vol. 7, pp. 22-34, Jan., 1995.
14. N. Virag, "Single Channel Speech Enhancement Based on Masking Property of the Human Auditory System," *IEEE tran. on speech and audio processing*, vol. **7**, pp. 126-137, Mar., 1999.
15. V. Stahl, A. Fischer and R.Bippus, "Quantile Based Noise Estimation for Spectral subtraction and Wiener Filtering" in ICASSP, 2000.
16. H. G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in Proc. ICASSP, pp. 153-157, 1995.
17. D. Kincaid, W. Cheney, Numerical Analysis, Brooks/Cole Publishing Company, 1996.
18. S. Quakenbush, T. Barnwell and M. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, 1988.
19. T. Painter and A. Spanias, "A Review of Algorithms for Perceptual Coding of Digital Audio Signals," *Digital Signal Processing Proceedings*, vol **1**, pp 179-208, 1997.
20. B. Scharf, Critical Bands, in foundation of modern auditory theory, New York : Academic Press, 1970.
21. T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proc. IEEE*, vol. **88**, pp. 451-513, 2000.
22. S. Quakenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, NJ:Prentice-hall, 1988.

저자 약력

● 윤 석 현 (Sukhyun Yoon)

2000년 2월: 한국과학기술원 전기 및 전자공학과 학사
2000년 3월~현재: 한국과학기술원 전기 및 전자공학과 석사과정
※ 주관심분야: 잡음처리

● 유 창 동 (Chang D. Yoo)

1986년: California Institute of Technology 학사
1988년: Cotnell University 석사
1996년: MIT 박사
1997년 1월~1999년 3월: 한국통신 Senior Researcher
1999년 3월~현재: 한국과학기술원 전기 및 전자공학과 조교수
※ 주관심분야: 음성신호처리, 디지털 신호처리