한국음향학회지 제21권 제4호 pp. 374~379 (2002)

유성음/무성음 분리를 이용한 잡음처리

Speech Enhancement Based on Voice/Unvoice Classification

유 창 동*
(Chang Dong Yoo*)

*한국과학기술원 전기 및 전자공학과

(접수일자: 2001년 12월 13일; 채택일자: 2002년 4월 9일)

본 논문에서는 유성음/무성음 분리를 이용하여 잡음처리를 한다. 유성음과 무성음은 음성의 하나의 중요한 특징으로 유성음과 무성음 부분에 각각 같은 잡음처리기법을 쓰는 것이 아니라 각각의 성질을 고려하여 잡음처리를 하였다. 유성음/무성음의 분리는 영교차율과 에너지를 이용하여 구해졌으며, 유성음/무성음 분리정보를 토대로 하여 변형된 음성/잡음우세결정방법을 제안하였다. 제안된 방법은 백색 잡음과 비행기 잡음에 오염된 음성문장에 대해 성능평가가 이루어졌다. 그리고 다양한 입력 신호대잡음비 (SNR)로 오염된 문장에 대해 세그멘탈 신호대잡음비를 구하고, 듣기 평가를 통해 기존의 방법보다 향상된 성능을 가짐을 알 수 있다.

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

In this paper, a novel method to reduce noise using voice/unvoice classification is proposed. Voice and unvoice are an important feature of speech and the proposed method processes noisy speech differently for each voice/unvoice part. Speech is classified into voice/unvoice using zero-crossing rate and energy, and a modified speech/noise dominant-decision is proposed based on voice/unvoice classification. The proposed method was tested on conditions of white noise and airplane noise, and on the basis of comparing segmental SNR with the existing method and listening to the enhanced speech, a performance of the proposed method was superior to that of the existing method.

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

I. 서 론

정보통신시대에 발맞추어 각종 음성처리 시스템이 발전하고 있으며, 그에 따라 고성능 음성잡음처리 시스템의 필요는 점점 증가하고 있다. 잡음이 있는 환경에서는 음성인식기의 인식률이 급격히 떨어지고, 이동 통신에서 원활하고 쾌적한 의사소통에 어려움을 겪게 된다. 그 동안 이러한 문제를 해결하기 위해 다양한 잡음처리방법이 제안되어 왔다: 주파수차감법에 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] 등이다.

최근에는 음성의 특징 중에 하나인 유성음/무성음의 분리를 통해 잡음처리를 하려는 움직임이 있다[15,16]. 유성음 부분은 빗살 여과기 (comb filter)에 통과시키고, 무성음 부분은 변형된 주파수 차감법을 이용하는 방법[15]과 유성음과 무성음을 각기 다른 음성모델을 적용하여 잡음처리하는 방법[16]이 나와 있다. 유성음과 무성음은 특성이 서로 다르기 때문에, 각 부분마다 같은 잡음처리기법을 쓰는 것이 아니라 각 특성에 따라 다르게 처리하여 좀더 잡음처리를 효과적으로 수행하는 것이 위의 방법의 목적이다. 특히 무성음은 에너지가 작아서 잡음처리 중에 잃

책임저자: 유창동 (cdyoo@ee.kaist.ac.kr)
305-701 대전시 유성구 구성동 373-1
한국과학기술원 전기 및 전자공학과
(전화: 042-869-3470; 팩스: 042-862-0559)

기 쉬운 부분이므로 이러한 과정을 통해 좀더 음성의 질(quality)과 명료도 (intelligibility)를 향상시키게 된다.

본 논문에서는 유/무성음분리 정보를 통해 더욱 향상된 음성/잡음 우세결정[17]을 수행한다. 유성음은 대부분의 음성성분이 저주파대역에 몰려 있고, 무성음의 경우는 고주파대역에 몰려 있다[18]. 이러한 특성을 이용하여 신호대잡음비가 낮은 상황에서도 음성성분을 최대한 보존하고 잡음을 제거하게 된다. 먼저 유/무성음의 분리는 영교차율(zero-crossing rate)과 에너지를 이용하여 분리를 한다. 영교차율은 배경잡음에 민감하게 반응하므로 미리 간단한 주파수차감법[1]을 통해 처리 (pre-noise canceling)한 후에 영교차율을 구한다. 그리고 각각의 프레임마다 에너지를 구하고, 앞서 구한 영교차율을 함께 이용하여 유/무성음을 분리해낸다. 유/무성음 분리 다음에는 유성음과 무성음에 따라 변형된 음성/잡음 우세결정이 이루어진다. 그 후에 잡음감쇄부를 거쳐 잡음처리를 하게 된다.

본 논문은 다음과 같이 이루어졌다. 제2장에서는 전체 잡음처리 시스템을 설명한다. 2.1절에서는 유/무성음을 어떻게 분리하는지 설명한다. 2.2절에서는 유/무성음 분리를 통해 변형된 음성/잡음우세결정을 소개한다. 제3장에서는 제안된 방법을 이용한 결과를 평가한다. 제4장에서는 내용을 재정리하고 결론을 내린다.

II. 잡음처리시스템

본 방법은 다음 세 가지 과정을 거친다. 첫째, 유/무성음 분리기에 통과시킨다. 먼저 배경잡음의 영향을 줄이기 위해 간단한 잡음처리를 거친다. 그리고 영교차율과 에너지를 구하고 그것을 바탕으로 유/무성음을 분리해낸다. 둘째, 유/무성음 분리를 바탕으로 변형된 음성/잡음우세결정을 한다. 셋째, 음성/잡음우세 결정에 의해 구해진 잡음추정치를 바탕으로 잡음을 처리한다. 그림 1은 전체잡음처리 시스템을 보여준다.

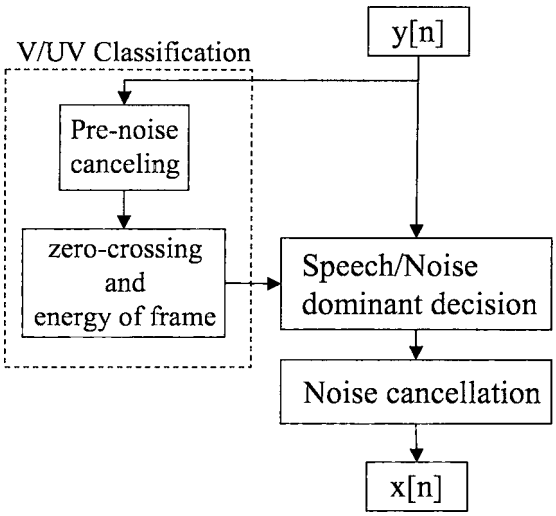


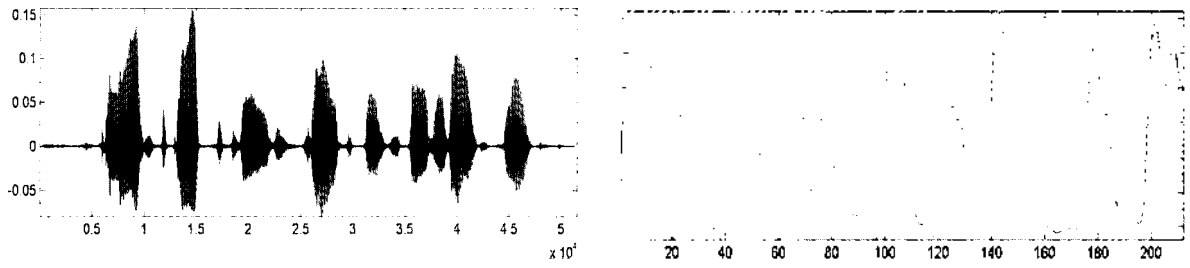
그림 1. 전체시스템
Fig. 1. The overall system.

2.1. 유/무성음의 분리

2.1.1. 영교차율 (zero-crossing rate)

유성음 성분은 저주파대역에 집중되어 있으며 무성음 성분은 고주파대역에 집중되어 있기 때문에, 유성음일 때 영교차율은 낮은 값을 가지게 되고 무성음일 때 영교차율은 높은 값을 가지게 된다[18]. 하지만 영교차율은 배경잡음에 민감하게 영향을 받기 때문에 영교차율을 구하기 전에 잡음처리 (pre-noise canceling)를 해주어야 한다. 간단한 주파수차감법[1]을 이용하여 미리 잡음처리를 해준 후에 영교차율을 구했다. 영교차율의 정의는 다음과 같다.

$$Z_m = \frac{1}{N}\sum_{n=1}^{N-1}\frac{|sgn(s_m[n]) - sgn(s_m[n-1])|}{2} \qquad (1)$$



(a) 간단한 전처리를 거친 후의 음성파형
(a) Speech waveform after pre-noise canceling

(b) 영교차율
(b) Zero-crossing rate

그림 2. 간단한 전처리를 거친 음성 파형을 통한 영교차율의 계산
Fig. 2. Calculation of zero-crossing rate using pre-noise canceled speech.

단, Z_m은 m번째 프레임의 영교차율이고 $s_m[n]$은 m번째 프레임에서의 n번째 샘플을 나타낸다. 그리고 $sgn(s_m[n]) = \begin{matrix} 1 \\ 0 \end{matrix}$ $\begin{matrix} \text{if } s_m[n] \geq 0 \\ \text{if } s_m[n] < 0 \end{matrix}$ 이다.

그림 2에서 (a)는 간단한 전처리를 거친 후의 음성의 파형, (b)는 영교차율을 보여준다.

2.1.2. 에너지

유성음과 무성음의 판정은 음성구간 내에서만 이루어져야 한다. 에너지를 통해 음성구간을 찾아낸다. 간단한 잡음처리 (pre-noise canceling)를 거친 후 일지라도 음성에는 약간의 잡음이 남아 있다. 이러한 남아있는 잡음까지 고려하여 음성구간을 찾아낸다.

에너지를 이용한 음성검출방법은 다음과 같다.

$$VAD_m = \begin{matrix} 1 \\ 0 \end{matrix} \quad \begin{matrix} \text{if } E_m \geq k * N_{m-1} \\ \text{if } E_m < k * N_{m-1} \end{matrix} \tag{2}$$

$$E_m = \sum_{n=0}^{N-1} s_m[n]^2 \tag{3}$$

$$N_m = \begin{matrix} N_{m-1} \\ p * N_{m-1} + (1-p) * E_m \end{matrix} \quad \begin{matrix} \text{if } VAD = 1 \\ \text{if } VAD = 0 \end{matrix} \tag{4}$$

$VAD_m = 1$일 때 음성구간을 나타내며, $VAD_m = 0$일 때 묵음구간을 나타낸다. E_m은 m번째 프레임의 에너지를 나타내며, N_m은 지금까지 업데이트된 잡음의 에너지를 나타낸다.

그림 3에서 (a)는 음성구간을 검출한 모습을 보여주며, (b)에서는 에너지 및 문턱치를 나타내었다.

2.1.3. 유/무성음의 분리

에너지를 통해 음성구간을 검출했고, 영교차율을 이용하여 유성음과 무성음을 분리하게 된다. 유성음과 무성음의 판정은 다음과 같이 한다.

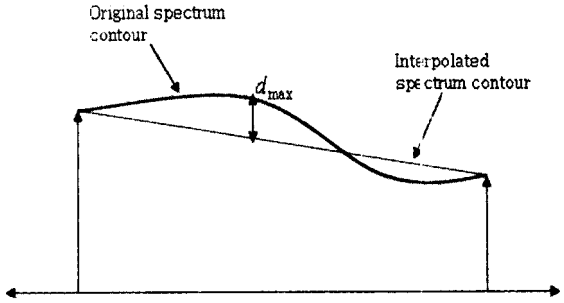


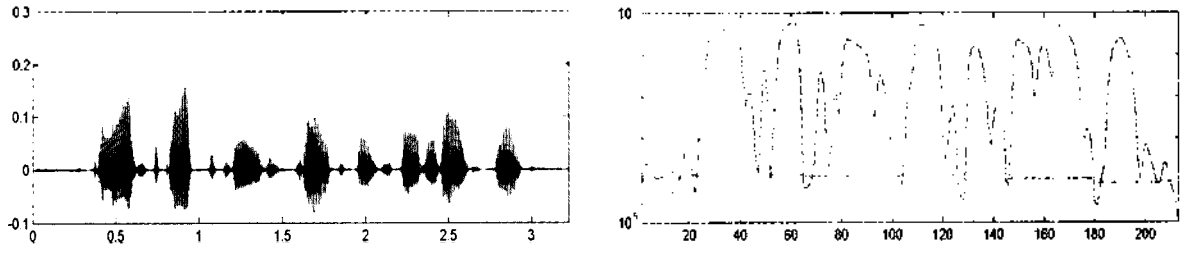
그림 4. 유성음과 무성음의 분리
Fig. 4. The classification of voice/unvoice.

```
If  VAD_m = 1
    If Z_m > N_unvoice
       s_m is unvoice.
    elseif Z_m < N_voice
       s_m is voice.                    (5)
    else
       s_m is unknown.
    end
end
```

문턱값들을 비교적 차이를 많이 두어 영교차율이 매우 낮을 때 유성음으로 판정하고, 매우 높을 때 무성음으로 판정한다. 그리고 영교차율이 $N_{unvoice}$와 N_{voice}의 사이에 있을 때는 판정이 모호한 부분으로 이때는 음성/우세 결정에서 특별한 고려사항이 없이 기존의 방법대로 처리해 나간다. 이것은 명확한 유성음의 성질과 무성음의 성질이 나타날 때에만 처리해 주고자 하는 것이다. 그림 4는 유성음과 무성음을 분리한 결과를 나타낸다.

2.2. 음성/잡음 우세결정

대역이 넓은 신호를 다룰 때는 고주파 부분의 음성신호가 매우 미약하여 고주파 부분의 부분적인 신호대잡음비 (local SNR)가 무척 낮게 나타난다. 신호대잡음비가 낮아졌을 때는 음성성분을 잡음과 구분해 내기가 무척 어렵다. 본 연구에서는 유성음/무성음이라는 음향학적인 성



(a) 음성구간을 검출한 모습
(a) Speech activity detection

(b) 에너지 및 문턱치
(b) Energy and threshold

그림 3. 에너지를 이용한 음성구간 검출
Fig. 3. Energy based speech activity detection.

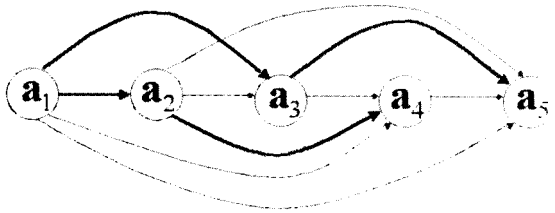


그림 5. 음성/잡음 우세결정 과정
Fig. 5. Speech/noise dominant decision process.

질을 이용함으로써 잡음처리의 효과를 좀더 높이는 것이 목적이다. 유성음은 대부분의 음성성분이 저주파대역에 몰려 있고, 무성음의 경우는 고주파대역에 몰려 있다[18]. 따라서 유성음으로 판정된다면 저주파부분의 초과추정(overestimate)부분을 완화시키고, 무성음으로 판정이 된다면 고주파부분의 초과추정 (overestimate)부분을 완화시킨다. 이는 잘못된 잡음우세결정을 내렸을 시에 음성성분을 최대한 보존하게 된다.

그림 5에서 음성/잡음 우세결정 과정을 보여 주며 다음 3가지 단계를 거친다[17]. 첫째, 잡음이 섞인 음성 $y[n]$을 $w[n]$으로 창을 씌운다. 창으로 씌워진 신호를 DFT (Discrete Fourier Transform) 계수로 변환한다. 각 임계밴드 안에 있는 계수의 크기들을 합한다. 그리고 과거 L개의 프레임 동안 각 임계밴드에서의 합들을 오름차순으로 정렬한다. 둘째, 근사함수를 이용하여 정렬된 수열을 모양에 따라 두 종류로 분류한다. 종류에 따라 다른 기준을 적용하여 유/무성음 판정을 이용하고 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. 구체적인 변형된 음성/잡음우세결정은 다음과 같다.

2.2.1. '종류1'에 대한 기준

'종류1'이란 최근 L개의 프레임 동안 i번째 임계밴드(critical band) 내에는 강한 음성성분이 있음을 나타낸다[17].

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

$(i,n)-region$는 잡음우세
$\Rightarrow |M[i,n]| = E[i, \lceil L \cdot high \rceil]/B_i$
$high \in [0.9, 1]$

만일 (고주파에서 $A[i,n] < \frac{1}{L}\sum_{q=1}^{L} E[i,q]$이고 무성음이면), **그때**

$\Rightarrow |M[i,n]| = E\left[i, \lceil L \cdot \frac{1}{2} high \rceil\right]/B_i$

그렇지 않으면,

$(i,n)-region$는 음성우세
$\Rightarrow |M[i,n]| = E[i, \lceil L \cdot low \rceil]/B_i$
$low \in [0.25, 0.35]$

단 $A[i,n]$: n번째 프레임의 i번째 임계 밴드 안에 있는 주파수 크기의 합.

$\{E[i,j]\}_{j=1}^{L}$: 모든 대역마다 길이 L인 수열 $\{A[i,j]\}_{j=n-L}^{n-1}$을 오름차순으로 정렬한 수열.

$M[i,n]$: n번째 프레임의 i번째 임계 밴드의 잡음 스펙트럼

B_i : i번째 임계 밴드 안에 있는 주파수성분(frequency bin)의 개수

$\lceil X \rceil$: 양의 무한대 방향으로 X와 가장 가까운 정수를 나타낸다. 예를 들어 $\lceil 1.2 \rceil = 2$.

2.2.2 '종류2'에 대한 기준

'종류2'란 최근 L프레임동안 i번째 임계 밴드 내에는 잡음이나 약한 음성성분이 대부분임을 나타낸다[17].

만일 ($A[i,n]$이 고주파에서의 값이면), **그때**
만일 무성음이 아니라면
$(i,n)-region$은 잡음우세
$\Rightarrow |M[i,n]| = c \cdot E[i,L]/B_i \qquad c \in [1,2]$
만일 무성음이라면
$(i,n)-region$은 음성우세
$\Rightarrow |M[i,n]| = E\left[i, \lceil \frac{1}{2} L \rceil\right]/B_i$

만일 ($A[i,n]$이 저주파에서의 값이면), **그때**
만일 유성음이 아니라면
$(i,n)-region$은 잡음우세
$\Rightarrow |M[i,n]| = c \cdot E[i,L]/B_i \qquad c \in [1,2]$
만일 유성음이라면
$(i,n)-region$은 음성우세
$\Rightarrow |M[i,n]| = E[i,L]/B_i$

잡음이 든 음성의 스펙트럼 $|Y[k,n]|$으로부터 잡음스펙트럼 $|M[i,n]|$을 빼어줌으로써 깨끗한 음성을 얻을 수 있다. 수학적인 표현은 다음과 같다.

$$S[k,n] = rect(|Y[k,n]| - |M[i,n]|) \quad k \in CB_i \quad (6)$$

단, $S[k,n]$은 잡음처리된 음성의 스펙트럼성분의 크기를 나타내고 $rect(\cdot)$는 반파정류기를 나타낸다. CB_i는 i번째 임계 밴드에 속하는 주파수 성분의 집합이다.

III. 평 가

제안된 방법은 세그멘탈 신호대잡음비 (segmental SNR)를 구하고 듣기평가 (MOS)를 실시하여 평가를 하였다. TIMIT 데이터베이스로부터 각각 남성과 여성에 의해 읽혀진 세 개의 문장으로 평가를 하였으며, 매개변수는 다음과 같이 선택되었다. 1)50% 오버랩을 하면서 길이 N=512인 해밍창 (Hamming window); 2)임계밴드의 개수는 22개; 3) a=0.3, c=2, $high$=0.9, low=0.3, k=1.5, p=0.8, N_{voice}=100, $N_{unvoice}$=180;

Noisex-92 데이터베이스에서 백색잡음과 비행기잡음을 사용하였다.

먼저 평가를 위해서 세그멘탈 신호대잡음비를 구하였다. 백색잡음와 비행기 잡음이 1 dB, 4 dB, 7 dB, 10 dB의 신호대잡음비로 섞여 있는 음성문장을 가지고 평가를 하였다. 그림 6과 그림 7은 각각 백색잡음과 비행기 잡음에 대해 입력 신호대잡음비대 향상된 세그멘탈 신호대잡음비를 나타낸다. 단 향상된 세그멘탈 신호대잡음비란 입력 세그멘탈 신호대잡음비와 처리 후의 출력 세그멘탈 신호대잡음비의 차이를 구한 값이다.

유/무성음 분리를 통한 방법이 유/무성음을 고려하지 않은 방법보다 더 나은 향상된 세그멘탈 신호대잡음비를 보임을 알 수 있다. 이는 유/무성음의 특성을 바탕으로 기존의 방법보다 음성성분을 강화시켰음을 보여준다. 또한

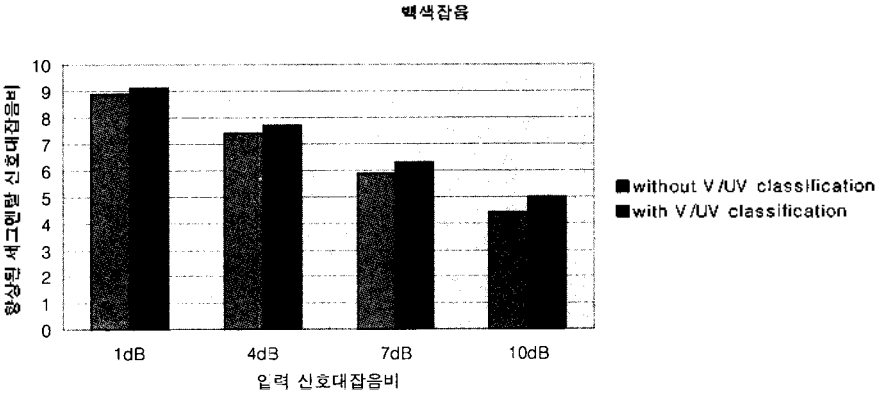


그림 6. 백색잡음에 대해 입력 신호대잡음비 대 향상된 세그멘탈 신호대잡음비를 나타낸 차트
Fig. 6. The chart of segmental SNR improvement versus initial SNR for white noise.

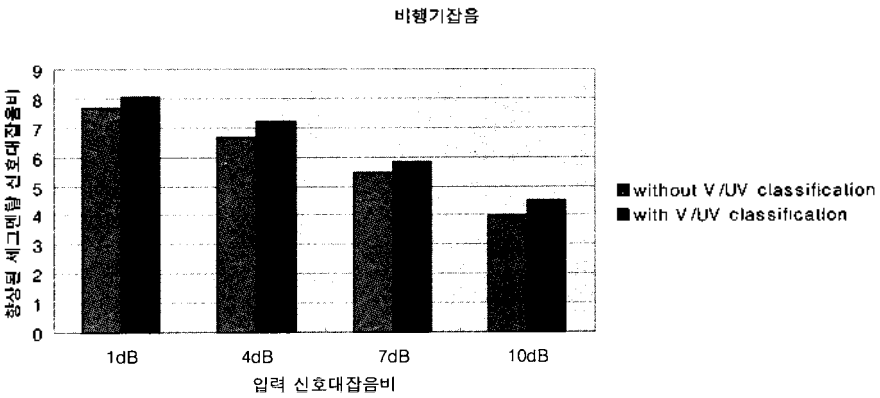


그림 7. 비행기잡음에 대해 입력 신호대잡음비 대 향상된 세그멘탈 신호대잡음비를 나타낸 차트
Fig. 7. The chart of segmental SNR improvement versus initial SNR for airplane noise.

표 1. 듣기평가 결과 (MOS)
Table 1. The result of listening test (MOS).

방법	백색잡음 (4dB)	백색잡음 (10dB)	비행기잡음 (4dB)	비행기잡음 (10dB)
기존의 방법[17]	3.12	3.43	2.94	3.25
제안된 방법	3.25	3.62	3.1	3.43

비정상적인 (non-stationary) 잡음인 비행기잡음에서도 유/무성음의 분리에 의해 더욱 향상되었음을 볼 수 있다.

본 논문에서는 객관적인 (objective) 평가인 세그멘탈 신호대잡음비뿐만 아니라 주관적 (subjective) 평가인 듣기평가[19]도 실시하였다. 백색잡음와 비행기 잡음이 4 dB와 10 dB의 신호대잡음비로 오염된 음성문장을 가지고 평가를 하였다. 다음 표 1은 백색잡음과 비행기잡음에 대한 듣기평가 결과를 나타내었다. 듣기평가에서도 제안된 방법이 기존의 방법보다 더 향상된 결과를 보여준다.

IV. 결 론

본 논문에서는 음성의 특징 중의 하나인 유성음/무성음을 분리하여 잡음처리를 하였다. 유성음과 무성음 부분에 각각 동일한 잡음처리기법을 쓰는 것이 아니라 각각의 성질을 고려하여 잡음처리를 하였다. 유성음/무성음의 분리는 영교차율과 에너지를 이용하여 구해졌으며, 유성음/무성음 분리정보를 토대로 변형된 음성/잡음우세결정방법을 제안하였다.

제안된 방법은 백색잡음과 비행기 잡음에 오염된 문장에 대해 평가가 이루어졌으며, 세그멘탈 신호대잡음비와 듣기평가를 통하여 기존의 방법보다 향상된 성능을 가짐을 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*, ASSP-27, 113-120, Apr. 1979.
2. J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, 67, 1586-1604, Dec. 1979.
3. M. Berouti and R. Schwartz, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE ICASPP*, Washington, DC, 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.*, 11, 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*, 28, 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*, 32, 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*, 33, 443-445, Apr. 1985.
8. Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Trans on Signal Processing*, 40, 725-735, Apr. 1992.
9. Y. Ephraim, "Statistical-model-based speech enhancement systems," *IEEE Proceeding*, 80, 1526-1555, Oct. 1992.
10. H. Sameti, H. Sheikhzadeh and L. Deng, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Tran. on Speech and Audio Processing*, 6, 445-455, Sep. 1998.
11. D. Tsoukalas, M. Paraskevas and J. Mourjopoulos, "Speech enhancement using psycho-acoustic criteria," *Proc. IEEE ICASSP*, 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*, 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*, 7, 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*, 7, 126-137, Mar. 1999.
15. D. O Shaughnessy and H. Tolba, Toward a robust/fast continuous speech recognition system using a voiced-unvoiced decision, *ICASSP*, 413-416, 1999.
16. Z. Goh, K. Tan and B. Tan, "Kalman-Filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. on Speech and Audio*, 510-524, Sep. 1999.
17. 윤석현, 유창동, "시간-주파수 영역에서 음성/잡음 우세 결정에 의한 새로운 잡음처리," 한국음향학회지 20 (3), 48-55, 4, 2001.
18. L. R. Rabiner, R. W. Schaferm, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
19. J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, Englewood Cliffs, NJ:Prentice-Hall, 1993.

저자 약력

● 유 창 동 (Chang Dong Yoo)
한국음향학회지 제20권 제3호 참조