논문 2008-45SP-2-11

상호억제와 시간지연 신경회로망을 사용한 적응적인 음성강조시스템

(An Adaptive Speech Enhancement System Using Lateral Inhibition and Time-Delay Neural Network)

최 재 승*

(Jae-Seung Choi)

요 약

본 논문에서는 다양한 배경잡음에 의해 열화된 음성을 강조하기 위하여 청각시스템을 기초로 한 적응적인 음성강조시스템을 제안한다. 제안한 시스템은 먼저 유성음과 무성음의 구간을 검출한 후, 각 입력 프레임에서 검출된 결과에 따라서 상호억제 계수와 진폭성분조정계수를 적응적으로 조정한다. 마지막으로 시간지연신경회로망을 사용하여 잡음신호를 제거한다. 실험결과, 본 시스템은 신호대잡음비의 평가방법을 통하여 다양한 잡음에 의해서 열화된 음성신호를 백색잡음 및 유색잡음에 대해서 효과적인 것을 보여준다.

Abstract

This paper proposes an adaptive speech enhancement system based on an auditory system to enhance speech that is degraded by various background noises. As such, the proposed system detects voiced and unvoiced sections, adaptively adjusts the coefficients for both the lateral inhibition and the amplitude component according to the detected sections for each input frame, then reduces the noise signal using a time-delay neural network. Based on measuring the signal-to-noise ratio, experiments confirm that the proposed system is effective for speech degraded by various noises.

Keywords: Speech enhancement, noise reduction, time-delay neural network, background noise, lateral inhibition.

I. 서 론

잡음이 존재하는 환경 하에서 음성인식 등의 음성정보처리의 실용화를 고려할 경우에 실제의 잡음환경 하에 대한 대응이 중요시되며 이에 대한 여러 연구가 다방면으로 진행되고 있다[1]. 이러한 연구 중에 잡음 하에서의 회화 및 음성인식에의 응용을 고려한 음성강조법이나 잡음제거의 방법으로 스펙트럼 차감법[2], 적응 필터법[3], 신경회로망(neural network, NN)에 의한 방법[4] 등 여러 방식이 발표되었다.

최근에 청각계를 기초로 한 모델[5~6] 및 시간지연신경회로망(time-delay neural network: TDNN)의 모델[7~8] 들을 계산기 상에 구축하려고 하는 연구가 다방면으로 진행되고 있다. 이러한 모델 중의 하나인 상호억제 및 TDNN이 본 연구에서 음성강조시스템으로 사용된다. 참고문헌 [5]는 오차역전파방식에 의해 학습된 3층 구조의 신경회로망을 사용하여 잡음량를 추정하였으며, 3종류의 상호억제 모델을 사용하여 각 프레임에서 잡음량을 적응적으로 추정하여 음성을 강조하였다. 참고문헌 [8]은 푸리에 변환의 진폭성분을 복원하는 잡음제거의 알고리즘을 사용하여 잡음이 중첩된 음성신호의 공간으로부터 잡음이 없는 음성신호의 공간으로 사상을 실행함으로써 특히 저역부의 잡음을 제거할 수 있었다.

음성신호는 시간변화가 중요한 정보 중의 하나이며,

* 정회원, 신라대학교 전자공학과
(Department of Electronics Engineering, Silla University)
접수일자: 2007년7월16일, 수정완료일: 2008년3월10일

NN을 이용하여 음성신호를 처리하는 경우도 시간구조를 NN에 구성하는 점도 중요하다. 따라서 본 논문에서는 NN에 시간요소를 도입한 시간이 지연된 TDNN[7~8]을 사용한다. 또한 음성신호를 고속 푸리에 변환(fast Fourier transform : FFT)한 경우 위상성분보다 진폭성분이 음성 정보를 많이 포함하고 있다. 따라서 본 논문은 스펙트럼 회복의 수단으로써 TDNN을 이용하여 FFT 진폭성분을 복원하는 알고리즘을 제안하며, 본 알고리즘을 사용하여 음성신호에 대한 잡음제거의 실험에 대한 유효성을 확인한다.

본 연구는 청각적 기강을 생리학적이 아닌 공학적으로 상호억제를 응용하려는 목적을 가지고 있다. 따라서 본 논문에서는 다양한 배경잡음환경 하에서 유효하는 청각시스템과 TDNN을 사용한 적응적 음성강조시스템을 제안한다. 먼저, 제안한 시스템은 잡음이 중첩된 음성신호를 캡스트럼 변환을 적용한 후, 이동평균에 의하여 구해진 스펙트럼 성분을 주파수영역에서 상호억제에 의해 컨볼루션한다. 그리고 제안한 시스템은 다른 경로에서 유성음과 무성음 구간을 검출한 후, 배경잡음을 제거하기 위하여 각 입력프레임에서 검출된 구간들에 따라서 상호억제계수와 진폭조정계수를 각각 조정한다. 마지막으로 잡음이 중첩된 음성신호는 제안된 TDNN을 사용하여 강조된다. 제안한 음성강조시스템을 평가하기 위하여, 음성의 명료도에 관계가 깊은 SNR을 사용하여, 백색잡음, 자동차잡음, 지하철잡음에 대해서 본 방식이 유효하다는 것을 명백히 한다.

Ⅱ. 음성신호 및 원 음성데이터

원 음성신호를 $s(t)$로 하였을 때 잡음이 부가된 음성신호 $x(t)$를 다음과 같이 나타낸다.

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

$n(t)$는 8kHz의 샘플링 주파수를 가진 백색잡음(white noise), 자동차잡음(car noise), 지하철잡음(subway noise)이다. 여기에서 백색잡음은 컴퓨터에 의해서 작성된 가우스 백색잡음이며, 자동차, 지하철잡음은 Aurora2 데이터베이스[9]에 포함된 잡음이다. 본 실험에서 사용한 음성 데이터는 8kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어숫자로 구성된 Aurora2 데이터베이스이다. 제안한 시스템은 Aurora2 데이터베이스로부터의 테스트 셋 A, B, C의 음성데이터와 테스트 셋 A의 자동차, 지하철잡음 등의 배경잡음을 사용하여 평가하였다. 본 실험에서 0dB에서 20dB까지의 다양한 신호대잡음비(Signal-to-Noise Ratio: SNR)가 부가된 잡음이 중첩된 음성신호를 사용하여 TDNN이 학습되었다. Aurora2 데이터베이스를 사용할 경우에 백색, 자동차, 지하철잡음을 Aurora2 데이터베이스의 음성신호에 부가한 후에 TDNN이 학습되었다.

Ⅲ. 시간지연신경회로망(TDNN)

FFT의 진폭성분은 위상성분보다 많은 음성정보를 포함하기 때문에, 본 논문에서는 스펙트럼 회복의 한 방법으로써 TDNN을 사용하여 FFT 진폭성분을 복귀하는 알고리즘을 제안한다. 본 논문에서는 어느 정도 비슷한 패턴마다 TDNN을 구축하는 경우가 학습하기 용이하다고 판단하여 유성부용 및 무성부용에 각각의 TDNN을 구축하는 그림 1의 개량된 TDNN 시스템을 제안한다[8].

먼저 잡음이 중첩된 입력음성신호 $x(t)$는 유성부 및 무성부로 판별된 후에 유성부 및 무성부에 따라서 각 프레임을 128샘플의 FFT진폭성분들로 분리한다. 그 후에 분리된 FFT 진폭성분들은 각각의 저역 및 고역의 TDNN에 부가된다. 각각의 저역용 TDNN 및 고역용 TDNN으로부터의 출력을 합성하여 최종 FFT 진폭성분을 구한다. 그러나 위상성분은 진폭성분으로부터 직접 구한다. 마지막으로 역 고속 푸리에 변환(inverse fast Fourier transform : IFFT)을 사용하여 강조된 음성신호 $y(t)$를 구한다.

제안한 TDNN은 역전파알고리즘을 사용하여 학습한

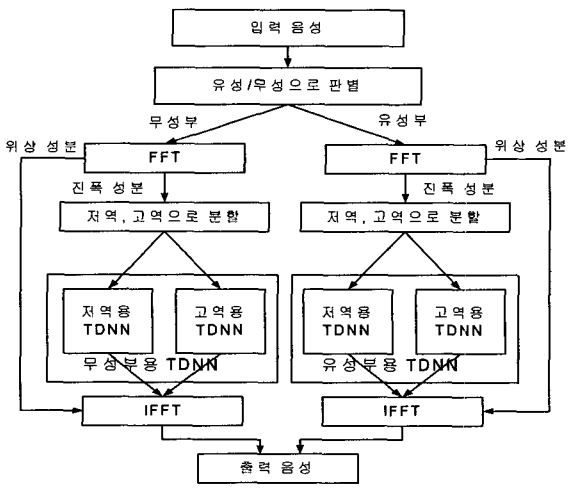

그림 1. 제안한 TDNN 시스템
Fig. 1. Proposed TDNN system.

다. 잡음이 중첩된 음성신호 $x(t)$로부터 구해진 FFT 진폭성분은 TDNN의 입력신호에 부가된다. 반면에 음성신호 $s(t)$로부터 구해진 FFT 진폭성분은 TDNN의 학습신호에 부가된다. 그 후에 TDNN은 각 프레임(128샘플)을 사용하여 학습된다. 단, TDNN에는 부가적인 정보로서 학습대상 프레임의 이전 2 프레임과 다음 1프레임이 부여된다. 따라서 입력되는 총 프레임 수는 4 프레임이다. 본 실험에서는, FFT에 의해 구해진 진폭성분은 FFT의 63번째의 성분을 중심으로 대칭한 값을 가지므로, 용장부를 제외한 0~63샘플이 저역 및 고역으로 분할되며, 각각의 결과들이 저역용 TDNN, 고역용 TDNN에의 입력으로 부가된다.

저역 및 고역주파수대역에 대한 2종류의 TDNN이 그림 2와 같이 구성된다. 32개의 FFT 진폭성분의 시간계열들은 n 프레임을 가진 입력층에 입력된다. 그 후에 입력층의 4 프레임은 첫 번째 중간층의 프레임에 연결된다. 64 유닛을 가진 첫 번째 중간층의 각 6프레임은 두 번째 중간층의 프레임에 연결된다. 그리고 32 유닛을 가진 두 번째 중간층의 각 프레임은 출력층에 연결된다. 본 실험에서 저역용 TDNN에의 입력신호는 FFT 진폭성분의 0~31샘플(0kHz~1.9kHz)이며, 여기에서 부가적인 입력신호로는 1개의 학습프레임과 2개의 이전프레임과 다음 1프레임으로 구성된다. 학습신호는 잡음을

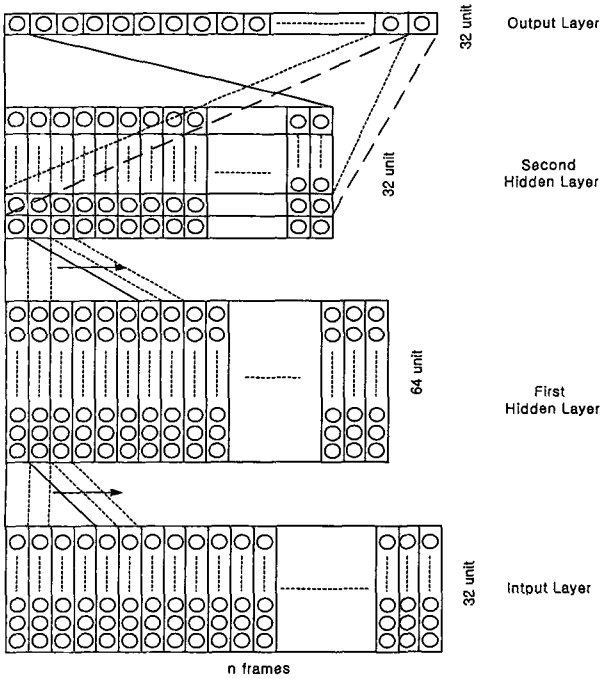

그림 2. 제안한 저역 및 고역용 TDNN의 구조
Fig. 2. Construction of proposed TDNNs for low and high frequency bands.

부가하지 않은 음성신호 중, 학습대상 프레임에 해당하는 프레임을 FFT함으로서 구해진 FFT 진폭성분의 0~31샘플이다. 반면에 고역용 TDNN의 입력신호는 고역용 TDNN에의 입력신호로서 부여된 FFT 진폭성분 32~63샘플(2kHz~3.9kHz)이며, 여기에서 입력신호는 위와 동일한 부가적인 프레임들이다. 학습신호는 잡음을 부가하지 않은 음성신호 중, 학습대상 프레임에 해당하는 프레임을 FFT하는 것에 의해서 구해진 FFT 진폭성분 32~63샘플이다. 본 실험에서는 제안한 TDNN들이 다음과 같은 5종류의 네트워크를 사용하여 학습되었다.

(1) $SNR_{IN}(Input\ SNR) = 20\,dB$,

(2) $SNR_{IN} = 15\,dB$,

(3) $SNR_{IN} = 10\,dB$,

(4) $SNR_{IN} = 5\,dB$,

(5) $SNR_{IN} = 0\,dB$.

학습의 실행에 필요한 각 TDNN의 학습조건으로는, 최대 학습횟수를 오차변화가 거의 없어지는 10,000회로 하였으며, 학습계수 α는 0.1, 가속도계수 β는 0.6, 초기 하중은 -0.06~0.06의 난수를 사용하였다.

Ⅳ. 음성의 특성 개선법

1. 이동 평균

프레임 사이의 단시간 스펙트럴의 창함수의 부엽에 의한 피크 및 잡음에 의한 불규칙적인 피크를 감소시켜서 명료한 음성을 구하는 하나의 방법으로 식 (2)와 같은 가중치가 부가된 이동평균을 제안한다[5].

$$\overline{P}(i,\omega) = \frac{1}{2M+1}\sum_{j=-M}^{M} W_j P(i-j,\omega) \qquad (2)$$

본 실험에서는 $M=2$로 하고, 가중치로는 $W_{-2} = W_2 = 0.7$, $W_{-1} = W_1 = 1.1$, $W_0 = 1.4$로 하였다. 여기에서 $\overline{P}(i,\omega)$는 평균화된 (i)번째의 프레임의 단시간 전력 스펙트럴이다.

2. 상호억제

FSLI(Function of Spatial Lateral Inhibition)는 내이의 기저막에 있어서 신경상호간의 상호억제 기강을 모의한 것이며, 음성의 스펙트럴의 높은 부분을 날카롭게 하며, 낮은 부분의 잡음을 억제하는 것으로부터 음성강조에 유효한 방법이다[5~6]. 따라서 FSLI의 모델은 그림

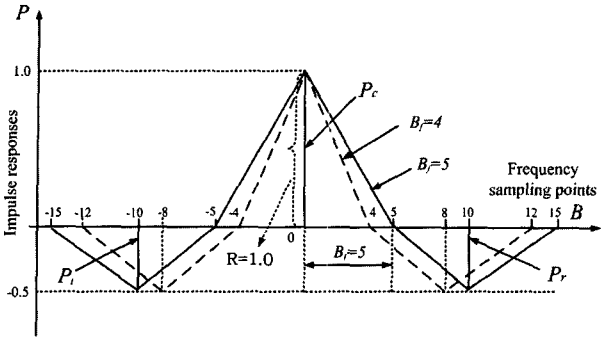

그림 3. 상호억제 함수의 임펄스 모델
Fig. 3. Impulse responses of lateral inhibition models.

3에 나타내는 것과 같은 1개의 흥분영역과 2개의 억제영역을 포함하는 주파수영역에서 사용된다. 그림 3은 본 실험에서 사용한 2종류의 FSLI의 특성을 나타낸 것이며, 가로 축은 주파수 표본점을 나타내고, 세로 축은 주파수 $B=0$의 위치에 단위입력이 부가된 경우에 그 근방의 표본점에서 얻어진 출력을 나타내고 있다. 그리고 B_f는 FSLI의 넓이를 결정하는 요소이며, R은 FSLI의 진폭을 결정하는 요소이다.

그림 3에서 FSLI의 진폭을 나타내는 요소 $P_{j(j=l,c,r)}$에 대하여 식 (3)과 같은 제한을 설정한다.

$$P_l + P_c + P_r = 0 \tag{3}$$

식 (3)의 제한은 상호억제에 의해 잡음의 합의 평균치가 영으로 되어서 잡음이 경감 된다. 본 실험에서는 식 (4)와 같은 값을 사용하여 상호억제를 하였다.

$$P_c = 1 \text{ and } P_l = P_r = -0.5 \tag{4}$$

V. 적응적 음성강조시스템

본 연구에 사용한 적응적음성강조 시스템을 그림 4에 나타낸다. 먼저, 8kHz로 샘플링된 잡음이 중첩된 음성신호 $x(t)$는 128샘플을 가진 프레임으로 분리되며, 해밍창 $W_1(t)$를 통과한 후에 캡스트럼변환을 한다. 또 다른 창 $W_2(t)$를 통과한 후, 저역에 해당하는 0번째부터 9번째까지의 캡스트럼 성분들이 구해지며, 음성신호에 대한 스펙트럴 성분들이 FFT에 의해서 구해진다. 3프레임 분의 지연이 일어난 후, 프레임 단위로 가중치를 부가하여 이동평균을 취한다. 다음에 이 스펙트럴 성분을 주파수 공간에서 FSLI를 한다. FSLI에 의해서 나타난 음의 성분들은 정류기에 의해서 제거되며, 이것을 진폭성분으로 한다(위의 경로). 한편, 다른 경로(하

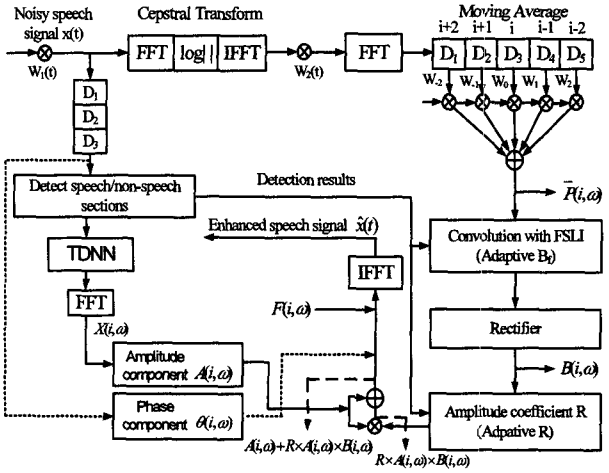

그림 4. 적응적 음성강조 시스템
Fig. 4. Proposed adaptive speech enhancement system.

단의 경로)에서의 잡음이 중첩된 음성신호 $x(t)$는 3프레임 분이 지연되며, 유성부와 무성부를 판별한 후에 제안한 TDNN에 의해서 잡음이 제거된다.

스펙트럼 차감법으로 알려진 J. S. LIM[2]의 방식에서는 잡음이 중첩된 신호로부터 음성신호의 단시간 스펙트럴진폭을 추정하기 위하여 4개의 매개변수 "a"(2.0, 1.0, 0.5, 0.25)가 4종류의 SNR(-5dB, 0dB, 5dB, ∞)에 따라서 최적인 값으로 선택됨으로써 음성의 명료도가 개선되고 있다. 본 연구의 경우에도 SNR_{IN}에 따라서 최적인 상호억제계수 B_f와 진폭성분조정계수 R의 값이 존재한다[5]. 그러므로 유성부 및 무성부로 판별된 결과들에 따라서 B_f와 R은 최적인 값으로 조정된다. 진폭성분 $A(i,\omega)$와 위상성분 $\theta(i,\omega)$가 구해진 후, 강조된 음성신호 $\hat{x}(t)$가 식 (5)를 사용하여 IFFT에 의하여 최종적으로 재생된다.

$$F(i,\omega) = A(i,\omega)(1 + R\times B(i,\omega))e^{j\theta(i,\omega)} \tag{5}$$

여기에서, i 및 ω는 각각 프레임번호 및 스펙트럴번호를 나타낸다.

일반적으로, 단시간에너지에 대한 문턱값은 각 잡음구간의 시작점에서부터 계산된다. 본 실험에서는 처음의 약 5프레임에서 각 문장의 평균 실효값 R_m을 구하여, 이 실효값의 $R_m/3$값이 문턱값 T_h가 되도록 실험적으로 정하였다. 즉, 각 프레임에서 $R_f > T_h$일 때에는 이 프레임은 유성부로 판별되며, $R_f < T_h$일 때에는 이 프레임은 무성부로 판별된다. 여기에서 R_f는 각 프레임에서 구해진 실효값을 나타낸다.

본 논문에서는 유성부 및 무성부를 판별함에 따른

오류에 대한 방법은 고려하지 않았지만, 그림 9에 나타내는 MOS(mean opinion score) 테스트에 의한 결과로부터 제안한 음성강조시스템은 MMSE-LSA의 결과와 비교하여 양호함을 확인할 수 있었다. 또한 비록 유성음 및 무성음 판별의 정확도에 따라서 약간의 성능개선은 기대할 수도 있겠지만, SNR에 의한 비교실험 및 MOS 테스트에서 보여주듯이 본 시스템의 전체 성능에는 영향을 주지 않을 거라고 판단되어진다. 그러나 유성음 및 무성음 판별오류에 따른 전체 성능 개선에 대한 알고리즘의 개발은 향후의 연구과제로서 검토가 필요하다.

VI. 실험결과 및 고찰

지금까지 기술한 기본조건 들을 사용하여, 제안한 시스템이 다양한 잡음에 의해서 열화된 음성에 대해서 SNR의 측정방법을 사용하여 유효하다는 것을 실험적으로 확인한다.

1. 진폭성분 조정계수 R에 대한 효과

본 시스템의 비교를 위하여, Aurora2 데이터베이스의 테스트셋 C로부터 임의적으로 20개의 문장이 선택되었다. 그림 5과 6은 자동차잡음이 부가되었을 때에 각 B_f에 대하여 R의 값을 조정함으로써 구해진 SNR_{OUT} (Output SNR)의 평균값들을 나타낸다. 그림 5와 6의 SNR_{OUT}의 평가값들로부터, 각 SNR_{IN}에 대한 최적의 R의 값들이 R을 조정함으로써 구해진다. 예를 들면, 그림 5의 $SNR_{IN}=15\,dB$의 경우에, 최대의 SNR_{OUT}값은 약 24 dB이 되므로 최적인 R의 값은 1.0이다. 따라서 SNR_{OUT}값은 $R=0.0$일 때의 잡음량을 포함한 원음에 대한 SNR_{IN}값 15 dB과 비교하여 약 9 dB 개선되었다. 여기에서, $R=0.0$은 자동차잡음에 의하여 열화된 원음에 대한 $SNR_{IN}=SNR_{OUT}$인 값을 나타낸다. 그림에는 표시하지 않았지만 그 외의 B_f 에 대해서는 SNR_{OUT}의 효과가 거의 없어서 본 실험에서는 $B_f=4,5$만을 사용하였다. 이러한 B_f와 R을 음성의 각 프레임에서 선정하는 방법은(참고문헌 [5]의 V.3절 참조), NN이 잡음이 중첩된 신호로부터 각 프레임의 입력 SNR의 대, 중, 소의 잡음량 상태를 추정하여, 각 프레임마다 SNR이 다르기 때문에 최적인 상호억제 계수 B_f와 진폭조정계수 R을 각각의 추정된 비율에 따라서 프레임마다 최적으로 조정하여 음성을

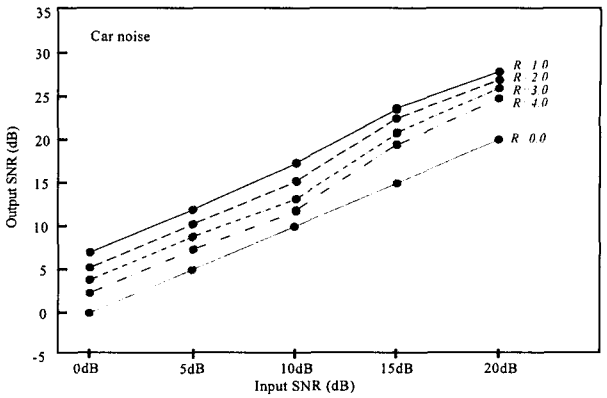

그림 5. R의 효과($B_f=4$인 경우)
Fig. 5. Effect of R (when $B_f=4$).

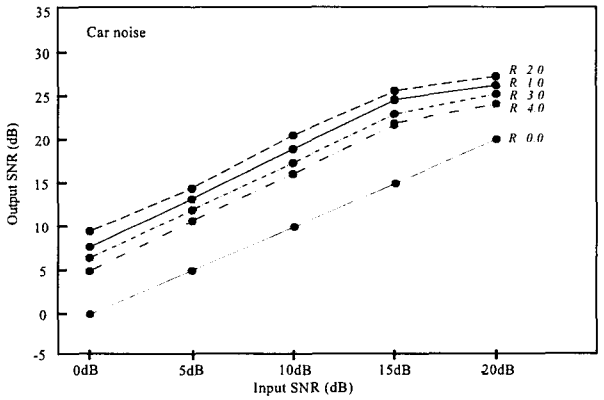

그림 6. R의 효과($B_f=5$인 경우)
Fig. 6. Effect of R (when $B_f=5$).

강조하게 된다.

그림 4의 제안한 적응적 음성강조 시스템을 사용할 때, 각 프레임에 대한 검출결과들이 무성부일 때에는 매개변수들이 $B_f=4$ 및 $R=1.0$으로 조정되며, 각 프레임에 대한 검출결과들이 유성부일 때에는 $B_f=5$ 및 $R=2.0$으로 조정된다.

2. 음성강조에 대한 성능평가 및 비교

본 절에서는 Aurora2 데이터베이스를 사용하여 여러 잡음환경 하에서 제안한 시스템에 대한 성능평가를 나타낸다. 본 시스템의 성능을 평가하기 위하여, Aurora2 데이터베이스의 테스트셋 A, B, C로부터 잡음이 중첩된 음성데이터들이 임의적으로 선택되었다. 제안한 시스템은 백색잡음, 자동차잡음, 지하철잡음 등에 대하여 MMSE-LSA(minimum mean-square error log-spectral amplitude)[10]와 비교되었다. 이 MMSE-LSA는 MMSE-STSA(minimum mean-square error short-time spectral amplitude)[11]를 기초로 하며, 통계적으로 독립적인 가우시안 랜덤 변수들을 사용함으로써 음성과

잡음의 스펙트럴 성분들을 유도한다. MMSE-LSA의 중요한 특징 중의 하나는 강조된 음성신호 속에서 "음악적 잡음"을 제거할 수 있다는 것이다. MMSE-LSA 방법을 실행할 때의 프레임길이는 128샘플(16ms)이며, 각 프레임에서 해밍창이 사용되었으며 중첩의 길이는 64샘플(8ms) 단위이다.

그림 7과 8은 백색잡음과 자동차잡음에 대하여 다양한 잡음레벨들($SNR_{IN} = 20dB \sim -5dB$)을 사용하여, 제안한 시스템과 MMSE-LSA 방법을 비교하여 30개의 문장에 대한 SNR_{OUT}의 평균값을 나타내었다. 그림 7의 백색잡음에 대하여, $R = 0.0$과 비교하였을 때, TDNN 만을 사용하였을 경우의 SNR_{OUT}의 최대 개선값은 약 4 dB, MMSE-LSA 방법의 SNR_{OUT}의

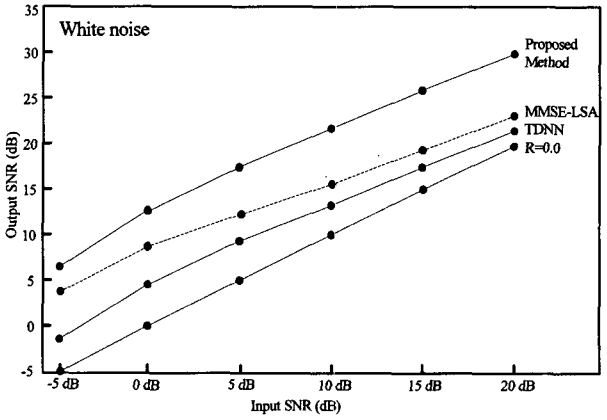

그림 7. 백색잡음 부가 시의 제안한 시스템과 MMSE-LSA 및 TDNN과의 비교

Fig. 7. Experimental comparison of proposed system and MMSE-LSA and TDNN methods when adding white noise.

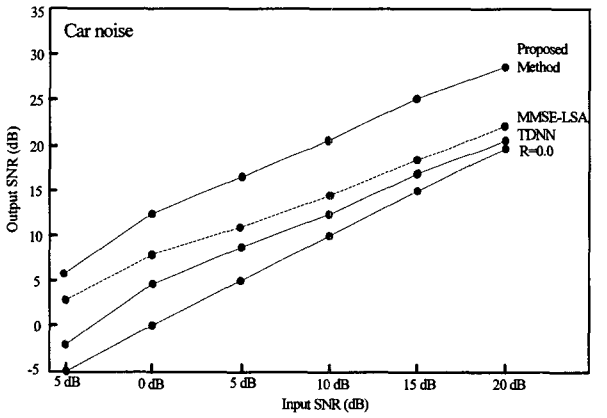

그림 8. 자동차잡음 부가 시의 제안한 시스템과 MMSE-LSA 및 TDNN과의 비교

Fig. 8. Experimental comparison of proposed system and MMSE-LSA and TDNN methods when adding car noise.

최대 개선값은 약 8 dB, 본 방법은 약 13 dB 개선되었다. 그리고 그림 8의 자동차잡음에 대해서도 같은 경향이 보여져, $R = 0.0$과 비교하였을 때, TDNN 만을 사용하였을 경우의 SNR_{OUT}의 최대 개선값은 약 3.5 dB, MMSE-LSA 방법의 SNR_{OUT}의 최대 개선값은 약 7 dB, 본 방법은 약 12 dB 개선되었다. 이상의 결과로부터, 사전에 TDNN에 의해 학습되지 않았던 $SNR_{IN} = -5dB$의 입력음성신호에 대해서도 본 시스템이 충분히 SNR을 개선함으로써 본 시스템의 성능 개선을 확인할 수 있었다.

그림에는 표시하지 않았지만 지하철잡음에 대한 SNR_{OUT}값은 MMSE-LSA 방법과 본 방법을 적용하였을 때의 자동차잡음의 경우보다 약 1.5 dB~2 dB 낮아졌다. 또한, 그림들에 나타낸 것과 같이, 제안한 시스템은 잡음레벨이 낮았을 때보다 잡음레벨이 높았을 때에 양호한 개선결과를 보였다. 이상의 결과로부터, FSLI 및 TDNN을 기초로 한 제안한 적응적인 음성강조시스템은 여러 잡음에 대하여 유효하다는 것을 말할 수 있으며, 제안한 시스템을 사용하였을 때 배경잡음들이 상당히 제거될 수 있다는 것을 의미한다.

음성품질의 측정으로 널리 알려진 MOS 테스트를 사용하여 본 시스템을 평가하였다. 제안한 시스템의 성능은 1에서 5까지의 MOS 점수를 사용하여 4명의 청취자(이중 1명은 영어를 모국어로 사용하는 사람)에 의해서 테스트되었다. 그림 9는 백색잡음과 자동차잡음에 대하여 입력 SNR이 0 dB, 5 dB, 10 dB인 경우에, 15종류의 테스트 음성에 대한 평균값을 나타내는 MOS 결과이다.

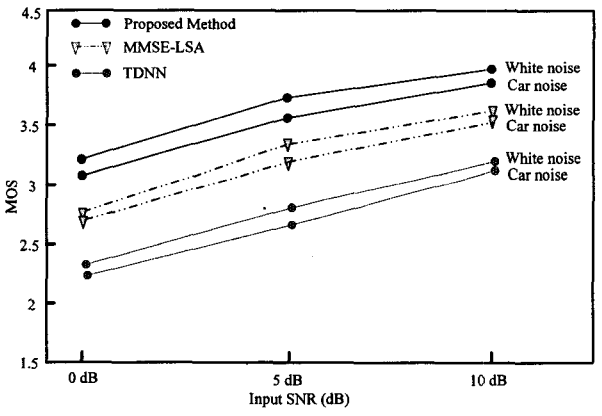

그림 9. 백색잡음과 자동차잡음 부가 시의 MOS 테스트에 의한 제안한 시스템과 MMSE-LSA 및 TDNN과의 비교

Fig. 9. The MOS values for the proposed method when compared with the MMSE-LSA and TDNN methods.

그림 9에 나타낸 바와 같이, 제안한 시스템의 MOS의 최대 개선값은 MMSE-LSA 방법과 비교하였을 때, 백색잡음의 경우에는 약 0.55 개선되었으며, 자동차잡음의 경우에는 0.40 개선되었다. 이상의 결과로부터, MOS 테스트에 의한 결과는 제안한 시스템이 MMSE-LSA 방법보다 양호하다는 것을 알 수 있으며, 본 시스템의 유효성을 확인할 수 있었다.

그림 10, 11, 12, 13은 자동차잡음이 부가된 경우의 남성화자의 음성신호 "4398515"에 대한 파형들을 표시한 것이다. 그림 10은 잡음이 부가되지 않은 경우의 입력음성 파형이고, 그림 11은 $SNR_{IN} = 5\,dB$인 경우에

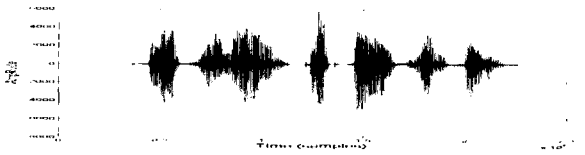

그림 10. 잡음이 없는 입력음성 신호의 파형
Fig. 10. The input of the clean speech signal.

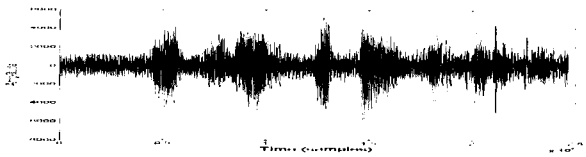

그림 11. 자동차잡음이 부가된 입력음성 파형 ($SNR_{IN} = 5\,dB$)
Fig. 11. The input of the contaminated speech signal, with car noise (in the case of $SNR_{IN} = 5\,dB$).

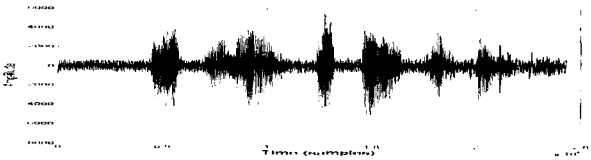

그림 12. TDNN만을 사용한 경우의 출력음성신호의 파형 ($SNR_{IN} = 5\,dB$)
Fig. 12. The waveform of the output speech signal using only TDNN (in the case of $SNR_{IN} = 5\,dB$).

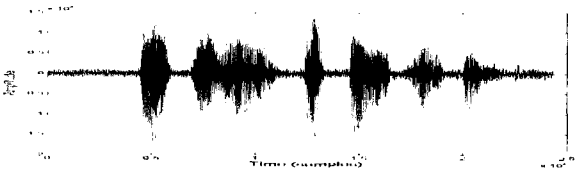

그림 13. 본 시스템에 의한 출력음성신호의 파형 ($SNR_{IN} = 5\,dB$)
Fig. 13. The waveform of the output speech signal using proposed method (in the case of $SNR_{IN} = 5\,dB$).

자동차잡음이 부가된 입력음성파형이다. 그림 12는 TDNN만 사용할 경우의 출력음성파형을 나타내고, 그림 13은 본 시스템에 의해 강조된 출력음성 파형이다. 따라서 그림의 파형에서 나타낸 것과 같이 본 시스템을 사용함에 의해서 배경잡음의 제거 및 음성강조의 모양을 알 수 있다.

Ⅶ. 결 론

배경잡음을 제거하기 위하여, FSLI 및 TDNN을 사용한 적응적 음성강조시스템을 제안하여, 본 시스템이 백색잡음, 자동차잡음, 지하철잡음에 대해서 유효하다는 것을 SNR을 사용하여 실험적으로 검증하였다. 따라서 제안한 적응적인 음성강조시스템은 프레임마다 추정되어진 최적한 값인 상호억제계수와 진폭조정계수를 조정함으로써 다양한 잡음이 중첩된 음성신호에 대하여 $SNR_{IN} = -5\,dB$까지 제거할 수 있었다.

향후의 연구과제로서는 본 논문에서는 각 입력프레임에 대해서 유성부와 무성부에 따라서 선택되어지는 2종류의 FSLI 만을 사용하였지만, 실용성을 고려하기 위하여 더 많은 FSLI를 사용할 것인가에 대한 연구가 필요하다. 이상으로, 본 연구에서 제안한 FSLI 및 TDNN을 사용한 음성강조시스템의 성과는 다양한 잡음 하에서의 음성강조에 도움이 될 것으로 생각된다.

참 고 문 헌

[1] K. K. Paliwal, "Neural net classifiers for robust speech recognition under noisy environments", IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp. 429-432, 1990.

[2] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise", IEEE Trans. Acoust., Speech, Signal Processing. Vol. 6, No. 5, pp. 471-472, 1978.

[3] B. Widrow, R. John, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, R. C. Goodlin, "Adaptive noise cancelling: Principles and applications", Proc. IEEE, Vol. 63, No. 12, pp. 1692-1716, 1975.

[4] W. G. Knecht, M. E. Schenkel, G. S. Moschytz, "Neural network filters for speech enhancement", IEEE Trans. Speech and Audio Processing, Vol. 3, No. 6, pp. 433-438, 1995.

[5] 최재승, "신경 회로망을 사용한 잡음이 중첩된 음성강조", 대한전자공학회 논문지, 제42권 5호 SP 편, pp. 165-172, 2005. 9.
[6] Y.M. Cheng, D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence". IEEE Trans. Signal Processing. Vol. 39, No.9, pp. 19431954, 1991.
[7] J.B. Hampshire, A.H. Waibel, "A novel objective function for improved phoneme recognition using time delay neural networks", IEEE Transactions on Neural Networks, Vol. 1, No. 2, pp. 216-228, 1990.
[8] 최재승, "시간지연 신경회로망을 이용한 잡음제거 시스템", 대한전자공학회 논문지, 제42권 3호 SP 편, pp. 121-128, 2005. 5.
[9] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", in Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, 2000.
[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985.
[11] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 32, No. 6, pp. 1109-1121, 1984.

저 자 소 개

최 재 승(정회원)
1989년 조선대학교 전자공학과 학사졸업
1995년 일본 오사카시립대학 정보통신공학과 석사졸업
1999년 일본 오사카시립대학 정보통신공학과 박사졸업
2000년~2001년 일본 마쯔시타 전기산업주식회사 AVC사 연구원
2002년~2007 경북대학교 디지털기술연구소 연구원, 프로젝트 리더
2007년~현재 신라대학교 전자공학과 교수
<주관심분야: 디지털통신, 음성신호처리, 신경회로망 등>