논문 2005-42SP-5-19

신경회로망을 사용한 잡음이 중첩된 음성 강조

(Speech Enhancement in Noisy Speech Using Neural Network)

최 재 승*

(Jae-Seung Choi)

요 약

잡음이 존재하는 환경 하에서 음성인식을 실시하는 경우, 잡음을 제거하고 음성을 강조하는 시스템이 필요하다. 따라서 우수한 스펙트럴 분석기강인 인간의 청각계를 모의하는 것은 음성강조에 있어서 효과적이다. 이러한 것을 구현하는 하나의 방법으로서 상호억제라고 하는 청각기강을 적응적으로 사용하는 방법을 제안한다. 이것은 신경회로망에 의해서 잡음의 크기를 추정하여 각 프레임에 대해서 그 크기에 따라서 적응적으로 상호억제 계수와 진폭성분조정 계수를 조정함으로써 음성을 강조하는 방법이다. 스펙트럴왜곡율 척도의 평가로부터 백색잡음뿐만 아니라 유색잡음 및 자동차의 주행잡음에 대해서도 본 방식이 효과적이라는 것을 확인한다.

Abstract

In speech recognition under a noisy environment, it is necessary to construct a system which reduces the noise and enhances the speech. Then it is effective to imitate the human auditory system which has an excellent analytical spectrum mechanism for speech enhancement. Accordingly, this paper proposes an adaptive method using the auditory mechanism which is called lateral inhibition. This method first estimates the noise intensity by neural network, then adaptively adjusts both the coefficients of the lateral inhibition and the adjusting coefficient of amplitude component according to the noise intensity for each input frame. It is confirmed that the proposed method is effective for speech degraded by white noise, colored noise, and road noise based on the spectral distortion measurement.

Keywords: Speech enhancement, Noise reduction, Lateral inhibition, Neural network.

I. 서 론

근년, 음성인식 등의 음성정보처리의 실용화를 위해서는 실제 잡음환경에 대한 대응이 중요시되고 있으며 이에 따라 여러 연구가 다방면으로 검토되고 있다. 이러한 연구 중에 잡음 하에서의 회화 및 음성인식 혹은 보청기에의 응용을 고려한 음성강조법이나 잡음제거의 방법으로 스펙트럼 차감(spectral subtraction)법[1][2][3][4], 위너 필터(Wiener filter)법[5], 적응 필터법[6], 신경회로망(Neural Network, NN)에 의한 방법[7] 등 여러 방식이 발표되었다. 이러한 방식 중에 스펙트럼 차감법은 잡음의 강도에 따라서 적응적으로 신호처리를 하는 것이 신호의 품질개선에 효과적이라는 공통적인 특징이 있다. 예를 들면, J. S. LIM[1]에서는 잡음이 중첩된 신호로부터 음성신호의 단시간 스펙트럴진폭을 추정하기 위하여 4개의 매개변수 "a"(2.0, 1.0, 0.5, 0.25)가 3종류의 신호 대 잡음비(Signal-to-Noise Ratio: SNR)(-5dB, 0dB, 5dB)에 따라서 최적값으로 선택됨으로써 명료도가 개선되고 있으며, J. S. LIM 등[2]에서는 SNR에 따라서 필터의 길이를 선택함으로써 명료도가 개선되고 있다. Y. M. Cheng 등[3]에서는 Itakura-Saito의 측정법을 사용하여 신호의 SNR이 적은 곳(SNR<5dB)에서는 처리방식I이 왜곡측도를 적게 하고, 신호의 SNR이 큰 곳(SNR>5dB)에서는 처리방식II가 왜곡측도를 적게 한다. 여기에서, 처리방식I은 상호억제(Function of Spectral Lateral Inhibition, FSLI)에 의한 단시간 전력스펙트럼 평균의 컨볼루션을 나타내며, 처리방식II는 FSLI에 의한 단시간 대수전력스펙트럼평균의 컨볼루션을 나타낸다. 본 연구의 경우에도 SNR에 따라서 최적인 상호억제계수 B_f와 진폭성분조정계수 R의 값이 존재한다.

반면에 최근에 청각생리학의 발전에 따라서 청각계에 의한 신호처리기강이나 기능이 명백해지고 있으며

* 정회원, 일본 오사카시립대학교 정보통신공학과
(Department of Information and Communication Engineering, Osaka City University)
접수일자: 2005년3월8일, 수정완료일: 2005년6월1일

이 모델을 계산기상에 구축하려고 하는 연구가 다방면으로 연구가 진행되어 지고 있다[3][8]. 이와 같은 청각모델의 연구는 청각기강을 해명함으로써 생리학 분야뿐만 아니라 인간의 음향신호처리에 관련한 분야의 공학적 응용의 연구에도 공헌 가능할 것으로 생각된다. 일반적으로 잡음의 강도는 음성신호가 포함되어있지 않는 시간영역에서의 신호강도로부터 구하기 때문에 잡음을 포함하는 음성신호에서 비 음성구간을 검출하는 것은 간단하지 않다[9]. 따라서 본 논문에서는 위의 문제점을 해결하는 하나의 방법으로써 여러 종류의 잡음환경 하에서 음성에 포함되는 잡음량을 신경회로망에 의해서 잡음량을 추정하는 방식을 제안한다. 다음으로 제안한 신경회로망에 의한 잡음량 추정방식을 기초로 하여 인간의 청각계를 모의한 내이(inner ear)의 기저막에 의한 상호억제 기강의 모델을 3종류로 확장하여 이 모델이 잡음량에 따라서 적응적으로 작용하는 적응적 음성강조 시스템을 제안한다.

본 논문에서는 음성 사이에 존재하는 무음이 포함된 음성데이터를 사용하여 실제적인 응용을 위하여 잡음이 중첩된 입력의 모든 프레임(frame)의 음성신호를 문장단위로 실효치가 일정하도록 정규화하여 실험을 하였다. 본 논문에서는 SNR_{seg}(Segmental Signal-to-Noise Ratio)가 약 -12dB 정도까지의 악 조건하에서 실험을 하여 본 방식의 유효성을 증명한다. 즉, 본 시스템은 입력신호에 대한 실효치를 정규화함으로써 진폭이 다른 미지의 잡음량을 포함한 입력에 대해서도 잡음량의 추정을 실시하여 광범위한 음성과 잡음의 입력에 대해서도 적응할 수 있는 새로운 적응적 음성강조시스템을 구현한다. 본 시스템은 백색잡음, 유색잡음, 자동차의 주행잡음에 대해서 각 프레임 단위로 잡음량에 따라서 상호억제 계수 B_f와 진폭성분 조정계수 R을 적응적으로 조정함으로써 음성을 강조한다. 음성특성의 개선척도로서는 음성의 명료도에 관계가 깊은 스펙트럴왜곡율 척도(Spectral Distortion, SD)를 사용하였다. SD에 대해서는 백색잡음 뿐만 아니라 유색잡음, 자동차의 주행잡음에 대해서도 본 방식이 유효하다는 것을 명백히 한다.

Ⅱ. 신경 회로망

1. 신경 회로망의 구성

음성신호에 포함되는 잡음량의 추정은 입력층, 중간층, 출력층의 3층 구조로 된 퍼셉트론 (perceptron) 형의 NN을 사용하여, 오차역전파방식 (error backpropagation, EBP)에 의해 학습시켰다. 잡음량이 다른 3종류의 캡스트럼(cepstrum) 데이터를 입력층에 입력하고 입력 유닛(unit)으로부터 결합계수 W_{ji}를 개입시켜 중간층의 유닛에 입력한다. 중간층으로부터 출력층에는 결합계수 W_{kj}를 개입시킴으로서 접속되어 각 유닛의 출력값이 계산된다. 다음으로 입력된 학습패턴에 대하여 학습데이터와 출력과의 오차의 절대값의 합 E가 구해져 이 오차 E가 일정한 값보다 적게 되도록 일반적인 오차역전파방식에 의해 결합계수가 변경된다. 유닛의 입출력함수는 비선형(sigmoid) 함수를 사용하고 결합계수의 수정에는 가속도계수를 사용하는 모멘트법(moment method)을 채용했다.

본 연구에서 NN을 사용하는 이유는 진폭의 크기가 시간적으로 일정하지 않는 유색잡음 및 자동차의 주행잡음에 대해서도 잡음량의 추정이 가능하도록 할 뿐만 아니라 EBP법이라는 학습알고리즘이 확립되어있는 효율적 학습법이 존재하기 때문이다.

2. 신경회로망의 학습법

한 프레임을 128샘플로 하는 음성의 대수 스펙트럴은 역 푸리에변환(Inverse Fast Fourier Transform, IFFT)된 후에 구해진 저역부의 10개의 캡스트럼을 NN에의 입력으로 하여 3종류의 잡음량을 추정한다. NN의 학습데이터는 (1) 잡음이 없는($k=0$) 상태를 [1.0, -1.0, -1.0], (2) 잡음이 중정도의 ($k=3$) 상태를 [-1.0, 1.0, -1.0], (3) 잡음이 많은 ($k=6$) 상태를 [-1.0, -1.0, 1.0]으로 설정하였다. 여기에서 k는 잡음량을 표시하는 요소이다. NN의 구성으로는 입력층의 유닛수 10에 대하여 중간층의 유닛수를 10, 15, 20, 30, 40의 5종류의 구성으로 검토하였으며 본 실험에서는 중간층의 유닛수를 30으로 하였다. 학습의 실행에 필요한 요소 및 여러 조건을 표 1에 나타낸다. 본 실험에서는 최대 학습 횟수

표 1. NN의 학습시의 여러 조건
Table 1. Various conditions for training of NN.

초기 가중치	-0.05 ~ 0.05의 난수
학습 계수	α = 0.1
가속도 계수	β = 0.5
네트워크의 구성	10 - 30 - 3
최대 학습 횟수	10,000회
입력의 실효값	1.0

를 오차변화가 거의 없어지는 10,000회로 하였다.

Ⅲ. 실험 조건

1. 음성신호

원 음성신호를 $s(t)$로 하여 잡음이 부가된 음성신호를 $x_k(t)=s(t)+k\times n(t)$로 나타낸다. 여기에서 $n(t)$는 컴퓨터에 의해서 작성된 가우스 백색잡음 혹은 음성의 스펙트럴의 경사특성과 거의 유사한 주파수 분포를 가지고 있는 유색잡음, 교통량이 많은 도로에서 녹음한 자동차의 주행잡음이며 샘플링 주파수는 모두 8kHz이다. 유색잡음은 그림 1의 진폭특성을 가진 저역통과 필터에 백색잡음을 통과시킴으로써 생성하였다. 결과적으로 구해진 유색잡음과 자동차의 주행잡음의 스펙트럴은 음성의 스펙트럴과 상당히 유사한 주파수 분포를 하고 있기 때문에 NN이 잡음량 k를 추정하는 것은 쉽지 않다고 생각되어지지만 본 실험에 의해서 충분히 식별가능하다는 것을 Ⅳ.3절에 나타낸다.

실험에 사용된 잡음이 중첩되어 있지 않는 음성신호의 각 문장의 실효치의 평균을 1.0으로 한 경우에 $k=1$에 해당하는 백색잡음과 유색잡음의 실효치는 0.40이며 자동차의 주행잡음의 실효치는 0.49이다. 잡음이 중첩된 입력의 실효치를 1.0으로 할 경우 $k=3$에 대해서 백색잡음과 유색잡음의 SNR_{seg}는 거의 -3dB이며 자동차의 주행잡음의 SNR_{seg}는 거의 -5dB이다.

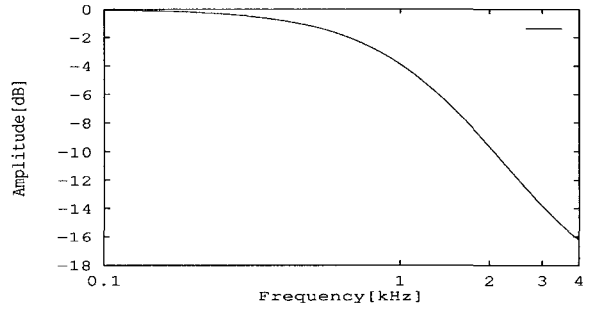

그림 1. 저역 통과 필터의 진폭 특성
Fig. 1. Amplitude characteristic of low-pass filter.

2. 원 음성데이터

음성데이터로서는 일본 음성정보처리 개발협회에서 배부한 연구용 연속음성 데이터베이스(DB)를 사용하였으며 총 26종류의 set(set1~set26)으로 구성된 본 DB는 일본 성인남성화자 48명 및 여성화자 54명이 발성한 총 22,000문장으로 구성되어 있다. 본 실험에서는 본 DB 중에서 무작위로 선택한 문장을 차단주파수 3.9kHz의 저역통과 필터를 통과시켜 8kHz로 샘플링한 데이터(부록 참조) 및 set1의 문장 중의 50문장을 실험데이터로 사용하였다. 표 2는 문장과 화자와의 관계를 나타내며 각 화자에 대한 문장 블록들은 서로 다른 화자이다. 또한 표2의 문장들은 전체 set 중에서 선택된 총 10개의 동일한 문장을 같은 화자가 발성한 데이터이다. NN의 학습데이터로서는 문장 F1에 백색잡음을, M1에 유색잡음을, M2에 자동차의 주행잡음을 부가해서 NN을 학습시켰다. 또한 학습에 사용하지 않은 문장은 학습 결과의 평가에 사용되었다. 표 3은 백색잡음, 유색잡음, 자동차의 주행잡음을 부가한 경우에 각각의 문장에 대하여 프레임 단위의 입력신호 대 잡음 비의 평균치 SNR_{seg}을 나타낸다[10]. 본 논문에 사용한 SNR_{seg}의 범위는 표 3, 5, 6에서 나타난 것과 같이 ∞~-12dB이다.

표 2. 사용한 문장과 화자와의 관계
Table 2. Relation between speech data and speakers.

남성 화자	문장	여성 화자	문장
화자 1	**M1,M2**	화자 1	**F1**
화자 2	M3	화자 2	F2

표 3. k와 SNR_{seg}과의 관계
Table 3. Relation between k and SNR_{seg} (dB).

문장	백색잡음 부가		유색잡음 부가		주행잡음 부가	
	$k=3$	$k=6$	$k=3$	$k=6$	$k=3$	$k=6$
M1	-3.07	-9.09	-3.23	-9.25	-5.35	-11.37
M2	-3.15	-9.17	-3.32	-9.34	-5.44	-11.46
M3	-3.06	-9.08	-3.17	-9.19	-5.36	-11.38
F1	-3.17	-9.19	-3.26	-9.29	-5.45	-11.53
F2	-3.10	-9.12	-3.32	-9.34	-5.51	-11.54

Ⅳ. NN에 의한 잡음량의 추정

1. 잡음량 추정시스템의 구성

잡음량 추정의 실험 시스템의 구성을 그림 2에 나타낸다.k에 해당하는 잡음량을 음성에 부가해서 이산시간신호 $x_k(t)$을 합성하여 한 프레임이 128샘플로 구성되는 해밍창 $W_1(t)$를 통과 시킨 후에 캡스트럼 변환을 한다. 구해진 캡스트럼을 방형창 $W_2(t)$에 통과시켜 저역부분의 직류성분을 포함하는 0번째부터 9번째까지의

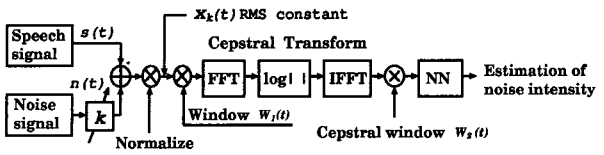

그림 2. 추정 시스템의 구성(NNES)
Fig. 2. Schematic diagram of estimating system (NNES).

10개의 캡스트럼 성분을 NN에의 입력으로 한다. $x_k(t)$는 전체 문장의 실효치가 일정하게 되도록 정규화 하였다. 이 경우, $k=6$의 잡음을 부가한 음성의 실효치를 정규화의 기준으로 하여 전체의 문장의 실효치를 이 값에 맞추었다.

2. 잡음량 추정율

NN의 학습이 종료된 시점의 결합계수를 사용하여 잡음량 k의 추정을 하였다. 잡음량 추정시스템의 성능 평가의 척도로서 "잡음량의 추정율"이라는 척도를 도입하였다. 이 추정율의 정의는 피 추정문장 입력으로 사용한 모든 프레임 수에 대하여 잡음량이 정확하게 추정된 프레임수의 비율이며 식 (1)과 같이 정의한다.

$$\text{잡음량의 추정율 (\%)} = \frac{\text{잡음량이 정확하게 추정된 프레임수}}{\text{입력에 사용된 프레임수}} \times 100 \quad (1)$$

3. 잡음량 추정의 실험

II.2절의 실험 조건 하에서 백색잡음, 유색잡음, 자동차의 주행잡음이 부가된 $k=0,3,6$의 3종류의 음성데이터를 사용하여 NN을 학습시켜 식 (1)의 잡음량의 추정율을 구하였다. 표 4는 화자와 문장이 학습문장과 다르더라도 각 잡음에 대하여 평균 98% 이상의 추정율이 구해지는 것을 나타낸다. 표에는 나타나 있지 않지만 학습문장과 추정문장이 동일한 경우, 백색잡음에 대해서는 평균 99% 이상의 높은 추정율이 구해졌으며 유색잡음 및 주행잡음에 대해서도 평균 98% 이상의 추정율이 구해졌다. 표의 결과로부터 유색잡음 및 주행잡음과 같이 음성신호의 스펙트럴에 유사한 잡음에 대해서도 NN에 의한 잡음량의 추정법이 유효하다는 것을 알 수 있다.

본 실험결과의 객관성을 검증하기 위하여 set1의 문장에 대해서 각 잡음에 대한 잡음량의 추정실험을 실시하였다. 표 5의 여러 문장에 대한 실험결과로부터 평균

표 4. 화자와 문장이 다른 경우의 추정율
Table 4. Correct estimation rates when speakers and speech data were not the same.

A: 백색잡음이 부가된 경우(Addition of white noise)

학습문장	추정문장	백색잡음 부가		
		$k=0$	$k=3$	$k=6$
M1, M2, F1	M3	100%	99.4%	99.6%
	F2	100%	99.2%	99.4%

B: 유색잡음이 부가된 경우(Addition of colored noise)

학습문장	추정문장	유색잡음 부가		
		$k=0$	$k=3$	$k=6$
M1, M2, F1	M3	100%	99.2%	99.4%
	F2	100%	99.0%	99.3%

C: 주행잡음이 부가된 경우(Addition of road noise)

학습문장	추정문장	주행잡음 부가		
		$k=0$	$k=3$	$k=6$
M1, M2, F1	M3	100%	99.1%	99.3%
	F2	100%	98.7%	99.1%

표 5. 각 잡음에 대한 테스트 set1의 잡음량 추정 결과
Table 5. Noise estimation results for test set1.

학습문장	부가잡음	Set1의 문장에 대한 잡음량		
		$k=0$	$k=3$	$k=6$
M1, M2, F1	백색잡음	99.7%	99.1%	99.3%
	유색잡음	99.4%	98.7%	99.0%
	주행잡음	99.2%	98.3%	98.8%
	평균	99.4%	98.7%	99.0%

98%이상의 정확한 잡음량 추정율이 구해진 것을 알 수 있다. 따라서 NN을 사용한 효과적인 잡음량 추정율이 화자, 잡음의 종류 및 잡음량에 의존하지 않는 것을 확인할 수 있었다.

V. 적응적 음성강조 시스템

1. 적응적 음성강조 시스템의 구성

본 논문에 사용한 적응적 음성강조 시스템(ASES)의 구성을 그림 3에 나타낸다. 먼저 잡음이 중첩된 음성신호를 각 문장마다 실효치가 일정하게 되도록 정규화하여 128샘플의 프레임마다 캡스트럼 변환을 한다(위의 경로). 프레임 단위의 지연(D_i)을 통과 시킨 후에 프레임 단위로 가중치(W_i)를 부가하여 스펙트럴 평균을 취한다. 이 후 3프레임 분의 지연이 발생한다. 이 때 입력신호 $x_k(t)$는 NN의 학습 시의 정규화 실효치와 동일한 수준의 신호로 정규화 되는데, 이것은 크기가 다른 미지의 잡음량을 포함한 입력에 대해서도 정확하게

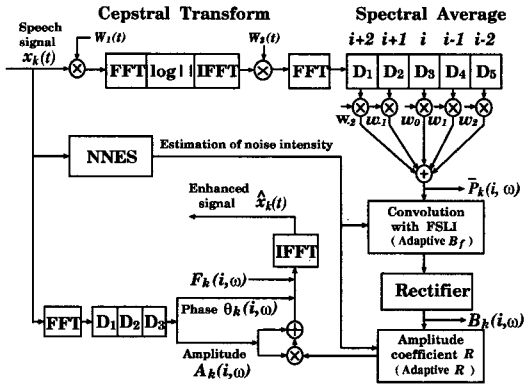

그림 3. 적응적 음성강조 시스템(ASES)
Fig. 3. Adaptive speech enhancement system(ASES).

NNES가 기능하도록 하기 위함이다. 다음으로 스펙트럴 평균으로부터 구해진 스펙트럴 성분을 주파수 공간에서 FSLI를 한다. 다만, FLSI에 의해서 생긴 부의 성분은 유용한 정보를 전달하지 않으므로 정류기(rectifier)에 의해서 영(zero)으로 처리된다. 그림 중의 $B_k(i,\omega)$는 FSLI된 후에 정류된 출력이다. NN의 학습 후에 IV.1절의 그림 2의 잡음량 추정시스템(NNES)이 각 프레임에서 잡음량의 크기에 따라서 최적인 B_f와 R을 조정하여 음성을 강조한다. 여기에서, NNES는 <V.3.나>절의 SD에 의한 R의 실험결과에 따라서 학습 후의 NN의 하중치에 의해서 R의 값을 각 프레임 단위로 잡음이 없는 상태를 $R=1.0$으로, 잡음이 중정도의 상태를 $R=2.0$으로, 잡음이 많은 상태를 $R=3.0$으로 조정한다. 한편, 다른 경로(하단의 경로)에서 푸리에변환(Fast Fourier Transform, FFT)되어 3 프레임분 지연된 후에 취해진 신호는 위상성분 $\theta_k(i,\omega)$와 진폭성분 $A_k(i,\omega)$로 분리되며 이러한 값들을 이용하여 식 (2)와 같이 $F_k(i,\omega)$가 구해진다. 따라서 $F_k(i,\omega)$ 신호를 IFFT함으로써 강조된 음성신호를 구할 수 있다.

$$F_k(i,\omega)=A_k(i,\omega)(1+R\times B_k(i,\omega))e^{j\theta_k(i,\omega)} \quad (2)$$

여기에서 i는 프레임 번호, ω는 스펙트럴의 번호를 나타낸다. 이상의 과정에 있어서 NNES의 출력인 잡음량에 적응시켜서 V.2절에서 기술하는 상호억제기강의 인펄스 모델에서의 상호억제 계수 B_f와 진폭성분조정계수 R을 최적한 값으로 조정하여 음성을 강조한다.

2. 스펙트럴 평균 및 상호억제

음성신호는 스펙트럴의 피크(peak) 성분에 매우 영향을 받기가 쉽다. 따라서 단시간 스펙트럴의 창함수(window function)의 부엽(sidelobe)에 의한 의외적인 피크로 인하여 강한 잡음환경 하의 음성강조에 있어서 음악적 잡음이 일어나지 않도록 하여야 한다[3]. 따라서 프레임 간의 잡음에 의한 불규칙적인 피크를 감소시켜서 명료한 음성을 구하는 하나의 방법으로 식 (3)과 같은 가중치가 부가된 스펙트럴 평균을 제안한다.

$$\bar{P}_k^{(i)}(i,\omega)=\frac{1}{2M+1}\sum_{j=-M}^{M}W_jP_k(i-j,\omega) \quad (3)$$

본 실험에서는 $M=2$이며, 가중치로는 $W_{-2}=W_2=0.7$, $W_{-1}=W_1=1.1$, $W_0=1.4$로 하였다. 여기에서 $\bar{P}_k^{(i)}(i,\omega)$는 평균화된 (i)번째 프레임의 단시간 전력 스펙트럴이다. 한편, FSLI는 내이의 귀저막의 신경상호간의 상호억제기강을 모의한 것이며 음성의 스펙트럴의 높은 부분(산과 같은 부분)을 날카롭게 하며 낮은 부분(계곡과 같은 부분)의 잡음을 억제시키는 효과가 있는 것으로부터 음성강조에 효과적인 방법이다[3][8]. 그림 4는 잡음량에 적응시켜 변화시키는 3종류의 FSLI의 인펄스 특성이다. 가로 축은 주파수 $B=0$에의 단위입력이 부가된 경우의 인펄스 응답을 나타낸다. 또한, B_f는 FLSI의 넓이를 정하는 요소이다. FLSI에 있어서 그림 4에 나타내는 것과 같은 인펄스 응답의 진폭을 나타내는 요소 $P_{j(j=l,c,r)}$에 대하여 식 (4)와 같은 제한을 설정한다. 이 제한에 의해 상호억제를 할 때 잡음의 합의 평균치가 영으로 되어 잡음의 경감이 이루어진다.

$$P_l+P_c+P_r=0 \quad (4)$$

본 실험에서는 $P_c=1$, $P_l=P_r=-0.5$로 하였다. 상

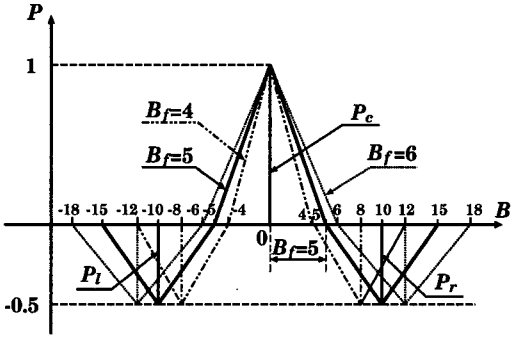

그림 4. 상호억제 기강의 인펄스 모델 응답
Fig. 4. Impluse responses of the lateral inhibition models.

호억제된 출력은 $\overline{P}_k^{(i)}(i,\omega)$와 그림 4에 나타내는 인펄스 응답과의 컨벌루션(convolution)에 의해 구해진다.

3. 잡음량에 적응적인 음성강조시스템의 실험

가. 잡음량에 대한 NN의 추정율

표 6은 학습이 종료된 음성강조시스템에서 피 추정문장 M3에 미지의 자동차의 주행잡음 k=0~7의 잡음량이 부가된 경우에 k=0, k=3, k=6으로 추정되는 추정율을 나타낸다. 표에서 나타나는 것처럼 자동차의 주행잡음에 대해서 $k=0,1$은 $k=0$으로, $k=2,3,4$는 k=3으로, k=5,6,7은 k=6과 같이 가장 높은 비율로 추정된다. 이 잡음량의 추정 결과를 그림 3의 음성강조시스템에 적용할 때, 예를 들면 부가된 자동차의 주행 잡음량이 k=4의 경우, 프레임마다 SNR_{seg}가 다르기 때문에, 프레임의 0.8%는 $k=0$으로, 98.0%는 k=3으로, 1.2%는 k=6으로 추정된다. 따라서 음성강조시스템에서는 각각의 추정율에 따라서 프레임마다 최적인 B_f와 R의 매개변수가 조정되어 음성을 강조하게 된다.

표 6. $k=0\sim7$에 대한 NN의 추정율(%)
Table 6. Estimation rates of NN for $k=0\sim7$.

학습 문장	주행 잡음량	SNR_{seg} (dB)	추정문장 M3		
			$k=0$	$k=3$	$k=6$
M1	$k=0$	∞	**100**	0.0	0.0
	$k=1$	4.18	**100**	0.0	0.0
	$k=2$	-1.84	1.3	**98.1**	0.6
M2	$k=3$	-5.36	0.2	**99.1**	0.7
	$k=4$	-7.86	0.8	**98.0**	1.2
F1	$k=5$	-9.80	0.2	0.8	**99.0**
	$k=6$	-11.38	0.0	0.7	**99.3**
	$k=7$	-12.72	0.0	0.0	**100**

나. 진폭성분 조정계수 R에 대한 효과

표 7은 추정문장 M3에 자동차의 주행잡음을 부가한 경우에 대해서 잡음량 k와 진폭성분조정계수 R을 매개변수로 하였을 때, 각각의 B_f에 대해서 R을 최적치로 조정하여 구한 출력 SD의 값과 입력 SNR_{seg}이다. 표 7의 SD의 평가치로부터 R을 조정함으로써 각각의 잡음량 k에 대해서 최적인 B_f 및 R의 값(dB)을 구할 수 있다. 예를 들면, 표 7.C의 k=7의 경우, 최소의 SD

표 7. SD에 의한 R의 효과(M3에 주행잡음을 부가)
Table 7. Effect of R by SD(Road noise is added to M3).

A: $B_f=4$의 경우(In the case of $B_f=4$)

잡음 강도	SNR_{seg} (dB)	진폭성분조정 계수 R				
		0.0	**1.0**	2.0	3.0	4.0
$k=1$	4.18	16.85	**11.78**	11.95	12.80	13.47
$k=2$	-1.84	19.45	**12.60**	12.84	13.32	13.86
$k=3$	-5.36	20.66	**13.98**	14.21	14.47	14.72
$k=4$	-7.86	21.36	**14.76**	14.88	15.06	15.32
$k=5$	-9.80	21.80	**15.38**	15.64	15.85	16.19
$k=6$	-11.38	22.12	**16.26**	16.67	16.94	17.13
$k=7$	-12.72	22.35	**16.20**	16.62	16.98	17.25

B: $B_f=5$의 경우(In the case of $B_f=5$)

잡음 강도	SNR_{seg} (dB)	진폭성분조정 계수 R				
		0.0	1.0	**2.0**	3.0	4.0
$k=1$	4.18	16.85	11.96	**11.80**	12.09	12.43
$k=2$	-1.84	19.45	12.02	**11.95**	12.15	12.62
$k=3$	-5.36	20.66	12.29	**12.11**	12.36	12.83
$k=4$	-7.86	21.36	12.46	**12.23**	12.52	12.97
$k=5$	-9.80	21.80	14.39	**14.15**	14.56	14.83
$k=6$	-11.38	22.12	16.12	**15.83**	16.33	16.52
$k=7$	-12.72	22.35	16.75	**15.36**	16.24	16.46

C: $B_f=6$의 경우(In the case of $B_f=6$)

잡음 강도	SNR_{seg} (dB)	진폭성분조정 계수 R				
		0.0	1.0	2.0	**3.0**	4.0
$k=1$	4.18	16.85	12.94	12.35	**12.02**	12.62
$k=2$	-1.84	19.45	13.72	13.15	**12.86**	13.54
$k=3$	-5.36	20.66	14.65	14.24	**14.09**	14.31
$k=4$	-7.86	21.36	14.89	14.50	**14.26**	14.67
$k=5$	-9.80	21.80	15.35	14.21	**13.02**	14.63
$k=6$	-11.38	22.12	14.42	13.06	**12.57**	13.58
$k=7$	-12.72	22.35	14.69	13.00	**12.43**	13.52

값은 12.43dB이 되므로 최적인 R의 값은 3.0이다. 따라서 이것은 R을 조정하지 않았을 때의 R=0.0에 해당하는 잡음량을 포함한 원음에 대한 SD의 값 22.35dB로부터 9.92dB 개선되어 있다. SD의 절대치가 다르기 때문에 단순히 비교하는 것은 어렵지만 전화의 PCM 방식의 평가 데이터에 의하면 20dB의 SNR의 개선량을 SD로 나타내면 약 3.5dB이라는 것을 생각하면(문헌[10]의 Fig. 3), 본 실험에서의 SD의 개선량은 최대 9.92dB이므로 상당히 유효하다는 것을 말할 수 있다. 실제로 강조된 음성을 들어본 결과 SD가 적으면 적을수록 양

호한 재생 음성이 구해졌다.

이상으로부터, 표 7은 본 실험에서 착목하고 있는 SD를 최소로 하는 의미에서 음성강조 매개변수 B_f 와 R에 최적치가 존재한다는 것을 명백하게 하고 있다. 그리고 본 실험 결과로부터 판단하면, 최적인 B_f와 R은 잡음량 k에 의존하지만 문장, 화자, 잡음의 스펙트럴 등에는 의존하지 않는°것으로 추측되어진다.

다. SD에 의한 적응적 음성강조 효과

그림 3에 나타내는 적응적 음성강조 시스템(ASES)에 있어서 NNES의 잡음량 추정 결과에 의해서, 어떤 프레임에 대해서 k=0으로 추정된 경우에는 B_f=4, R=1.0으로, k=3으로 추정된 경우에는 $B_f=5$, R=2.0으로, k=6으로 추정된 경우에는 $B_f=6$, R=3.0으로 되도록, 각각의 매개변수가 조정된다. 이러한 B_f와 R의 값은 표 7에서 구한 것과 동일한 방법으로 표 2의 문장에 대해서 구한 최적값이다. 그림 5, 6, 7은 시스템의 출력 $\hat{x}_k(t)$의 SD값을 나타낸 것이다. 그림 중의 "R=0.0"은 잡음량을 포함한 원음에 대한 SD값이고 "Optimal-R"은 $B_f=4$로 고정하여 R만을 최적치로 조정함으로써 구한 SD값이다. 또한, "ASES"는 V.1절에서 제안한 본 방식에 의해서 구해진 SD값이다. 즉, B_f와 R을 위의 최적치로 조정해서 구한 SD값이다. 그림 5의 백색잡음에 대한 SD의 평가를 보면, "R=0.0"에 대하여 "Optimal-R", "ASES"의 순서로 SD값이 각각 최대 약 7.36dB 및 12.86dB 개선되어져 있다. 그림 6의 유색잡음에 대해서는 "R=0.0"에 대해서 "Optimal-R", "ASES"는 각각 최대 약 7.23dB 및 11.58dB의 SD가 개선되었다. 그리고 그림 7의 주행잡음에 대해서도 비슷한 경향이 보여져, "R=0.0"에 대해서 "Optimal-R",

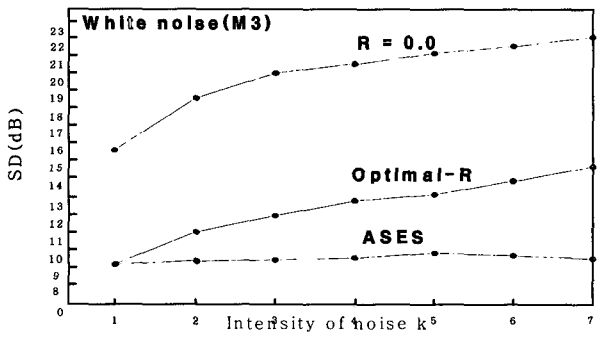

그림 5. SD로 측정한 백색잡음의 경우의 음성강조효과
Fig. 5. Effect of speech enhancement measured with SD when white noise is added.

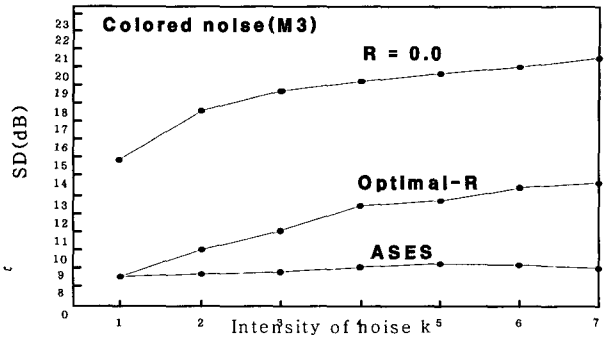

그림 6. SD로 측정한 유색잡음의 경우의 음성강조효과
Fig. 6. Effect of speech enhancement measured with SD when colored noise is added.

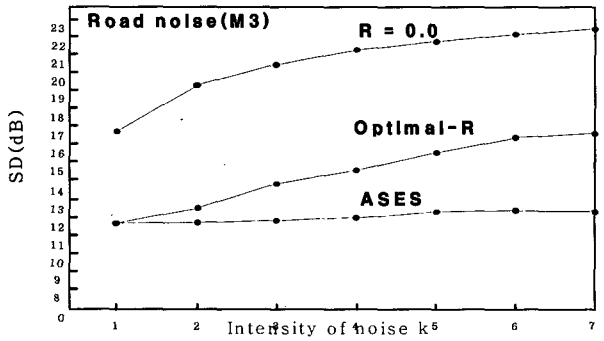

그림 7. SD로 측정한 주행잡음의 경우의 음성강조효과
Fig. 7. Effect of speech enhancement measured with SD when road noise is added.

"ASES"는 각각 최대 약 6.15dB 및 9.92dB의 SD가 개선되었다. 또한, 그림으로부터 알 수 있듯이 잡음량이 많아지면 SD의 개선량이 증가하는 경향이 있다. 따라서, 최적화에 의해서 백색잡음, 유색잡음, 주행잡음에 대해서 각각 5.51dB, 4.35dB, 3.77dB의 개선이 되었다. 잡음량 $k=1$의 경우에 있어서는 "Optimal-R"에 대한 "ASES"의 SD값은 거의 개선되어 있지 않으나 "$R=0.0$"에 대한 "ASES"의 SD의 전체 개선량은 다른 잡음량 k와 비슷하게 개선되어져 있는 것을 알 수 있다. 향후, $k=1$에 대한 잡음량의 개선방법의 검토를 고려 중에 있다. 이상의 결과로부터, 잡음량 추정에 의한 적응적 음성강조시스템이 백색잡음을 비롯하여 유색잡음 및 주행잡음에 대해서도 유효하다는 것을 말할 수 있다.

VI. 결 과

인간의 청각계에 있어서 중요한 역할을 하고 있는 상호억제기강의 모델을 잡음량에 따라 적응적으로 작용하게 하는 적응적 음성강조 시스템을 제안하여 본 시스템

이 백색잡음, 유색잡음, 자동차의 주행잡음에 대해서 유효하다는 것을 실험적으로 검증하였다. 이상, 본 실험으로부터 구해진 결과를 정리하면 다음과 같다.

(1) SNR_{seg}가 -12dB정도까지의 음성에 대해서 NN으로 양호한 잡음량의 추정이 가능하였다.

(2) 화자와 문장이 학습데이터와 다르더라도 평균 98% 이상으로 백색잡음, 유색잡음, 주행잡음에 대하여 잡음량의 추정이 가능하였다.

(3) 프레임마다 추정되어진 잡음량에 대해서 최적한 B_f와 R을 구하여 이 매개변수를 적응적 음성강조시스템에 사용함으로써 SD값을 개선 가능하게 하였다.

(4) 잡음억제의 효과는 유색잡음 및 주행잡음에 대해서도 크지만 특히 백색잡음에 대해서 현저하였다.

이상, 본 연구에서 제안한 FSLI를 사용한 적응적 음성강조시스템은 백색잡음뿐만 아니라 유색잡음 및 주행잡음에 대해서도 상당히 효과적인 음성강조시스템이라는 것을 실험적으로 검증하였다. 이 성과는 음성인식과 소음 하의 회화에 대한 음성강조에 도움이 될 것으로 생각된다.

부 록

음성데이터베이스로 사용한 일본어문장은 다음과 같다. 괄호 안의 문장 들은 일본어를 한국어로 번역한 것이다.

1. M1: "Kaijou wa dochiradesuka("회의장은 어디입니까?")
2. M2:"Kankoukyoukai deshouka"("관광협회입니까?")
3. M3:"Kondono onsei kenkyuukaio kikini ikitaindesukeredomo douittara iindeshouka("이번 음성 연구회에 참석하고 싶습니다만 어떻게 가야하나요?")
4. F1:"Donnatokoroga arimasuka("어떤 곳이 있습니까?")
5. F2:"Kanazawani ryokoushitainode sochiranokankouchinitsuite shitsumon shitaindesukeredo("가나자와에 여행을 가고 싶습니다만 그 곳의 관광지에 대해서 질문하고 싶습니다")

참 고 문 헌

[1] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-26, No.5, pp. 471-472, 1978.

[2] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," IEEE Trans. Acoust., Speech, Signal Processing, Vol.26, No.4, pp. 354-358, 1978.

[3] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," IEEE Trans. Signal Processing, Vol. 39, No. 9, pp. 1943-1953, 1991.

[4] S. F. BOLL, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Processing, Vol.27, No.2, pp. 113-120, 1979.

[5] T. V. Sreenivas and P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement," IEEE Trans. Speech and Audio Processing, Vol.4, No.5, pp. 383-389, 1996.

[6] B. Widrow et al., "Adaptive noise cancelling: Principles and applications," Proc. IEEE, Vol. 63, No. 12, pp. 1692-1716, 1975.

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

[8] S. A. Shamma, "Speech Processing in the Auditory System II: Lateral Inhibition and the Central Processing of Speech Evoked Activity in the Auditory Nerve", J. Acoust. Soc. Am. Vol.78, No.7, pp. 1622-1632, 1985.

[9] Y. Wu, Y. Li, "Robust speech/non-speech detection in adverse conditions using the fuzzy polarity correlation method", IEEE International Conference on Systems, Man, and Cybernetics, Oct. pp. 2935-2939, 2000.

[10] K. Itoh, N. Kitawaki, K. Kakehi, "A Study of Objective Quality Measures for Digital Speech Waveform Coding Systems", IEICE, Vol. J 66-A, No. 3, pp. 274-281, 1983.

저 자 소 개

최 재 승(정회원)

1989년 조선대학교 전자공학과 졸업(공학사)

1995년 일본 오사카시립대학 정보통신공학과(공학석사)

1999년 일본 오사카시립대학 정보통신공학과(공학박사)

2000년~2001년 일본 마쯔시타 전기산업주식회사 AVC사 연구원

2002년~현재 경북대학교 디지털기술연구소 연구원, 프로젝트 리더

<주관심분야: 음성 및 영상 신호처리, 잡음제거, 신경망, 영상 워터마킹, 디지털 TV 등>