논문 2005-42SP-3-16

음성강조에의 응용을 위한 신경회로망에 의한 잡음량의 추정법

(Estimation method of noise intensity by neural network for application in speech enhancement)

최 재 승*

(Jae-Seung Choi)

요 약

잡음이 중첩된 음성으로부터 잡음을 제거하기 위해서는, 잡음의 크기에 따라서 음성처리 시스템의 매개변수를 변경하는 것이 양호한 음질의 음성을 재생하는데 바람직하다. 본 논문은 백색잡음 및 자동차의 주행잡음에 의해 저하된 3단계의 음성을 학습할 수 있는 3층 구조의 신경회로망을 사용하여, 음성 중의 잡음량의 크기를 추정하는 방식을 제안한다. 실험결과, 제안한 방법은 신경회로망에 의해서 잡음량이 추정될 수 있는 것을 알 수 있었으며, 화자와 음성 데이터가 학습데이터와 다르더라도 백색잡음에 대해서 평균 95% 이상의 높은 잡음 추정율을 구할 수 있었다.

Abstract

To reduce the noise in the noisy speech, it is desirable to change the parameters of the speech processing system according to the noise intensity to reproduce a good quality speech. This paper proposes an estimation method of noise intensity using a three layered neural network, which is able to learn the three graded speeches that is degraded by white noise or road noise. Experimental results demonstrate that the noise intensity could be estimated by the neural network. Even if the speakers and speech data are different from the training data, estimation rates for the noise intensity can be estimated by the neural network with an average accuracy of 95% or more for white noise.

Keywords: Speech enhancement, Noise reduction, Neural network, Estimation method, Noise intensity.

I. 서 론

음성인식 시스템의 실용화를 실현하기 위해서는 잡음제거의 처리가 필요하다. 이러한 잡음 제거는 음성장치의 전 처리로서 필요할 뿐만 아니라 음성의 명료도를 증가시켜 청각적 피로도를 감소시키는 효과가 있다.

소음하의 회화나 음성인식 또는 보청기에의 응용을 고려한 음성강조 및 잡음제거를 위한 스펙트럼 차감법(spectral subtraction)[1][2][3][4][5], 위너필터(Wiener filter)[6], microphone array[7][8], 신경회로망[8][9], 적응 필터법[10][11] 등의 방식이 발표되었다. 이러한 논문 중에 스펙트럼 차감법은 잡음의 강도에 따라서 적응적으로 신호 처리를 하는 시스템이 필요하다는 공통된 특징을 가지고 있다. 예를 들면, J. S. LIM[1]에서는 매개변수(parameter) "a"가 신호 대 잡음비(Signal-to-Noise Ratio: SNR)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[4]에서는 Itakura-Saito의 측정법을 사용하여 신호의 SNR이 적은 곳에서는 처리방식 I이 왜곡측도를 적게 하고, 신호의 SNR이 큰 곳에서는 처리방식 II가 왜곡측도를 적게 한다. 본 연구의 경우에도 SNR에 따라서 최적인 진폭성분조정계수 R의 값이 존재한다[5]. 일반적으로 잡음의 강도는 음성신호가 포함되어있지 않는 시간영역에서의 신호강도로 부터 구하기 때문에 잡음을 포함하는 음성신호에서 비음성구간을 검출하는 것은 간단하지 않다[12].

본 연구는 장래적으로 유색잡음에 대해서도 잡음량

* 정회원, 일본 오사카시립대학교 정보통신공학과
(Department of Information and Communication Engineering, Osaka City University)
접수일자: 2004년11월29일, 수정완료일: 2005년2월4일

의 추정이 가능한 시스템을 구성하여, 비음성구간의 검출을 필요로 하지 않는 음성강조시스템 등에 쉽게 사용할 수 있는 잡음량 추정방식을 개발하는 목적으로 진행한 연구이다. 본 연구는 다양한 종류의 잡음환경 하에서 발성된 음성을 신경회로망(neural network, NN)에 학습시킴으로써 음성에 포함되는 잡음의 크기를 추정하는 시스템을 제안한다.

Ⅱ. 잡음량 추정 시스템의 개요

음성신호에 포함되는 잡음량의 추정에는 3층 구조의 퍼셉트론(perceptron)형의 NN을 사용하여 역전파(Back Propagation : BP) 알고리즘으로 학습시켰다. 이하의 실험 결과로 부터 알 수 있듯이 3층 구조의 NN으로도 충분히 본 연구의 목적을 달성할 수 있었다.

1. 신경회로망의 구성

본 연구에 사용한 NN은 그림 1에 나타난 것과 같이 3층 구조의 전결합형이다.

출력층과 중간층의 임의의 유닛(unit)은 고유의 가중치계수 w_i에 의해서 모든 하위층의 유닛과 결합되며, 각 유닛의 입력 및 출력관계는 식 (1)과 같은 비선형(sigmoid) 함수이다. 가중치의 초기치는 ±0.07의 범위내에 들어오도록 난수(random number)를 사용하였다.

$$f(x) = \frac{2.0}{1.0 + \exp\left(-\sum_i w_i x_i + \theta_i\right)} - 1.0 \qquad (1)$$

여기에서, x_i는 해당 유닛에의 입력이고, θ는 해당 유닛의 문턱값(threshold)이다.

2. 음성 신호

원 음성신호를 s(t)로 하면, 잡음이 중첩된 음성신호 $x_k(t)$를 식 (2)와 같이 나타낸다.

$$x_k(t) = s(t) + k \times n(t) \qquad (2)$$

여기에서, n(t)는 컴퓨터에 의해서 작성한 가우스(gauss) 백색잡음이고 샘플링 주파수는 8kHz이다. 그리고 k는 잡음강도를 나타내는 계수로 본 연구에서는 0, 3, 6의 값을 취한다. 각 단문의 실효치의 평균을 1.0으로 한 경우의 n(t)의 실효치는 0.16이며, 이 실효치는 k=6에 대한 SNR_{seg}이 거의 0dB이 되도록 n(t)의 실효치를 실험적으로 결정하였다. 본 연구에서 추정하는 SNR_{seg}의 범위는 표 5에서 알 수 있듯이 ∞~2dB이다.

3. 시스템의 구성

본 연구에서 제안한 시스템의 구성을 그림 2에 나타낸다. 샘플링 주파수 8kHz의 이산시간신호 $x_k(t)$를 방형창 $W_0(t)$에 의해 (A) 128 혹은 (B) 256 샘플(sample)의 프레임(frame)으로 분리하여 각 프레임의 실효치 R_f를 구하여, 이 값이 문턱값 $T_h = r_m/3$ 보다 적은 프레임만을 사용한다. 여기에서, r_m은 실험을 단순화하기 위하여 시스템에 추가되며 각 문장의 문장 전체에서 구한 실효치이며, 실제적 응용에서는 이동평균을 이용하는 등의 방법이 필요하다. 각 프레임의 샘플값을 해밍창(Hamming window) $W_1(t)$를 통과시킨 후, 켑스트럼변환(cepstral transform)을 한다. 구해진 켑스트럼을 방형창 $W_2(t)$에 통과시켜 단시간 영역성분을 추출한다. NN의 입력으로는 각 프레임의 (A), (B)의 샘플수에 대응해서 10개 혹은 20개의 켑스트럼을 사용한다.

4. 원 음성 데이터

본 실험에서 사용한 음성 데이터는 음성정보처리 개발협회에서 배부한 연구용 연속음성 데이터베이스 중에서 성인남성 화자 3명 및 성인여성 화자 3명으로 구성된 문장을 차단주파수 3.9kHz의 저역 통과 필터를 통과시켜 8kHz로 샘플링 하였다. 사용된 음성데이터는 12종

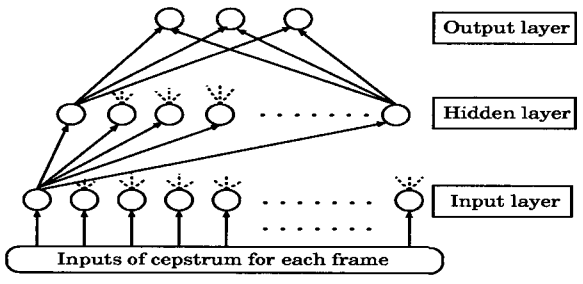

그림 1. 본 시스템에 사용한 3층 구조의 NN의 구성
Fig. 1. Construction of three layer NN system.

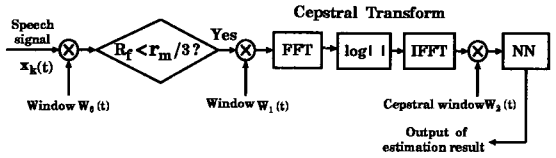

그림 2. 시스템의 구성
Fig. 2. Schematic diagram of the proposed system.

표 1. 사용한 문장과 화자와의 관계
Table 1. Relation between speech sentences and speakers.

남성화자	음원	여성화자	음원
화자 1	**M1, M2**	화자 1	F1, F2
화자 2	M3, M4	화자 2	**F3, F4**
화자 3	M5, M6	화자 3	F5, F6

류의 일본어 문장으로 구성된다(부록 참조). 표 1은 화자와 음원과의 관계를 나타낸다. NN에의 학습데이터로서는 음원 M1, M2, F3, F4를 사용하였으며, 나머지 음원은 학습결과의 추정에 사용하였다.

Ⅲ. 실 험

1. 실험조건과 추정법

음성의 대수 스펙트럴(spectral)을 역 푸리에 변환(Inverse Fourier Transform)하여 구한 켑스트럼은 음성추정에 유효한 매개변수이다. 음성의 스펙트럴 포락은 켑스트럼의 저 quefrency에 의해서 거의 완전하게 나타낼 수 있다. 이 때문에 켑스트럼 전체에서 10% 정도를 추출하면 충분하기 때문에, 본 실험에서는 1프레임을 128샘플로 하였을 경우에는 켑스트럼의 저역부의 10개를, 256샘플인 경우에는 20개의 켑스트럼을 NN에의 입력으로 하여 3종류의 잡음량을 추정하였다. NN에의 타겟(target) 신호는, (T1) 잡음이 없는 (k=0) 상태를 [1.0, -1.0, -1.0], (T2) 잡음이 많은 (k=3) 상태를 [-1.0, 1.0, -1.0], (T3) 잡음이 아주 많은 (k=6) 상태를 [-1.0, -1.0, 1.0]으로 설정하였다. NN의 구성은 표 2에 나타낸 것과 같이 2종류의 입력에 대하여 5종류의 중간층으로 하여 총 10종류를 검토하였다. 학습의 실행에 필요한 여러 조건을 표 3에 나타낸다. 본 실험에서는 최대 학습횟수를 오차변화가 거의 없어지는 10,000회로 하였다.

그림 3은 1프레임의 샘플수 N이 256일 경우에 대해서 중간층의 유닛수를 10, 15, 20, 30, 40으로 하였을 때의 NN의 학습시의 오차곡선을 나타낸다. 여기서 오차곡선은 하나의 네트워크에 대해서 10회의 시행에 대한 평균값이다. 그림에서와 같이 중간층의 유닛수가 30인 경우가 수속성이 가장 뛰어났기 때문에 본 실험에서는 중간층의 유닛수를 30으로 하였다.

본 실험에서는 잡음량 추정시스템의 성능평가의 척도를 위하여 식(3)의 잡음량 추정율을 도입한다.

표 2. NN의 구성
Table 2. Construction of NN.

입력층의 유닛수	중간층의 유닛수	출력층의 유닛수
10	10, 15, 20, 30, 40	3
20	10, 15, 20, 30, 40	3

표 3. NN의 학습시의 여러 조건
Table 3. Various conditions for training of NN.

초기 가중치	-0.05 ~ 0.05의 난수
학습 계수	α = 0.1
가속도 계수	β = 0.6
최대 학습횟수	10,000회
입력의 실효값	1.0

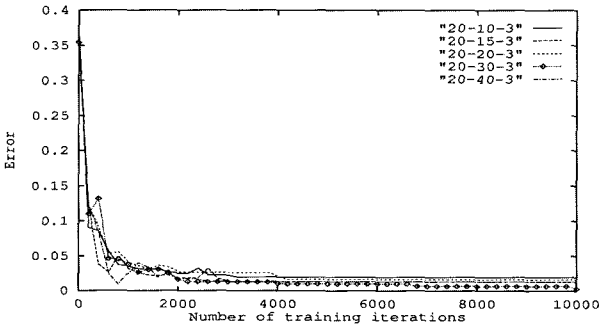

그림 3. 중간층의 유닛수에 대한 학습시의 오차곡선
Fig. 3. Error curves of training for the number of units in the hidden layer.

$$\text{잡음량의추정율}(\%) = \frac{\text{잡음량이정확하게추정된프레임수}}{\text{입력에사용된프레임수}} \times 100 \quad (3)$$

2. 학습 문장간의 거리 측정

패턴(pattern) 공간상에서 패턴의 유사성을 측정하기 위해서, 켑스트럼 영역에서의 각 음성간의 유클리드(Euclid) 거리를 측정하여 거리와 NN에 의한 잡음량의 추정성과의 관계를 명확하게 하였다. 학습에 사용한 3종류의 잡음량을 가진 음성 데이터의 켑스트럼을 $x_{k,j,i}$로 나타낸다. 여기에서 k는 각 잡음량 0, 3, 6을 나타내며, $j=1,2,3,\cdots,F$는 프레임 번호를, $i=1,2,3,\cdots,Cep$은 각 프레임에서의 켑스트럼의 quefrency의 순서를 각각 나타낸다. 본 실험에서의 Cep은 10 혹은 20이다. 켑스트럼의 중심을 식 (4)로 나타낸다.

$$C_{k,*,i} = \frac{1}{F}\sum_{j=1}^{F} x_{k,j,i} \quad (4)$$

각 프레임에서의 켑스트럼과 켑스트럼의 중심까지의 거리의 표준편차를 식 (5)로 나타낸다. 여기에서, $d_{k,j,*} = \sqrt{\sum_{i=1}^{Cep}(x_{k,j,i} - C_{k,*,i})^2}$, $m_k = \frac{1}{F}\sum_{j=1}^{F} d_{k,j,*}$ 이다. 또한 켑스트럼의 중심 간의 거리를 식 (6)으로 나타낸다.

$$\sigma_k = \sqrt{\frac{1}{F}\sum_{j=1}^{F}(d_{k,j,*} - m_k)^2} \quad (5)$$

$$d_{kk'} = \sqrt{\sum_{i=1}^{Cep}(C_{k,*,i} - C_{k',*,i})^2} \quad (6)$$

표 4. 학습 데이터간의 거리 $d_{kk'}$와 표준편차 σ_k ($T_h = r_m/3$의 경우)

Table 4. Distance $d_{kk'}$ and deviation σ_k of training data (in the case of $T_h = r_m/3$).

A: N=128, Cep=10의 경우(In the case of N=128, Cep=10)

음원	d_{03}	d_{06}	d_{36}	σ_0	σ_3	σ_6
M1	1.447	1.804	0.359	0.285	0.057	0.036
M2	1.620	1.978	0.360	0.301	0.062	0.039
F3	1.331	1.690	0.360	0.336	0.089	0.051
F4	1.241	1.589	0.349	0.334	0.086	0.047

B: N=256, Cep=20의 경우(In the case of N=256, Cep=20)

음원	d_{03}	d_{06}	d_{36}	σ_0	σ_3	σ_6
M1	2.827	3.533	0.714	0.559	0.108	0.064
M2	3.212	3.936	0.727	0.579	0.096	0.054
F3	2.665	3.384	0.721	0.680	0.184	0.104
F4	2.411	3.109	0.700	0.568	0.149	0.084

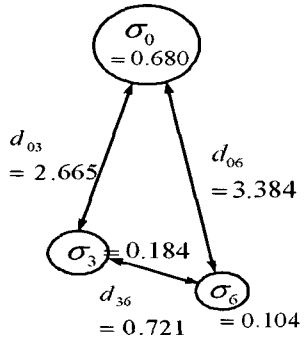

그림 4. 거리 및 표준편차 사이의 관계의 예

Fig. 4. An example of the relation between the distances and their deviations.

표 4는 켑스트럼의 중심 간의 거리 $d_{kk'}$와 표준편차 σ_k를 나타낸다. 표의 제1열은 음원의 종류를 나타내고, 각 행은 중심 간의 거리 및 거리의 표준편차를 나타낸다. 그림 4는 샘플수 N=256 및 음원 F3의 경우에 대한 켑스트럼 그룹의 중심 간의 거리와 표준편차를 각각 직선과 원으로 표시한 예이다. 그림으로부터 k=0의 켑스트럼 그룹(상측의 원 근방)과 k=3 및 k=6의 켑스트럼 그룹(하측의 왼쪽과 오른쪽 원 근방)은 상당히 떨어져 있지만, k=3과 k=6의 켑스트럼 그룹은 상당히 접근되어 있다. 또한 평균값을 중심으로 한 분포도는 σ_0, σ_3, σ_6의 순서로 적어지고 있다. 그림은 각 데이터간의 중첩이 적고 NN에 의한 추정이 가능하다는 것을 나타내고 있다.

Ⅳ. 잡음량 추정의 실험 결과

k=0, 3, 6의 3종류의 음성 데이터를 사용하여 NN을 학습시켰다. 학습에 사용한 음성 및 학습에 사용하지 않은 음성의 각각에 대해서, (1) 켑스트럼의 수 Cep, (2) 문턱값 T_h, (3) 문장, (4) 화자의 4가지의 매개변수를 변화시켜서 잡음량 추정 실험을 하였다. 그리고 문턱값은 시스템의 복잡함을 줄이기 위하여 Ⅱ.3절의 설명에서는 $T_h = r_m/3$과 같이 일정한 값으로 하였지만, 여기에서는 $T_h = r_m/5$ 및 $T_h = r_m$에 대해서도 실험을 하였다.

1. 음성 데이터에 대한 입력의 SNR_{seg}

표 5는 각각의 음원에 대하여 각 프레임의 신호 대 잡음비의 평균값 SNR_{seg}를 나타낸다[13].

2. N과 Cep 및 T_h와의 관계

문장 M1, M2, F3, F4로 NN을 학습시킨 후 추정 실험을 하였다. 표 6은 문턱값 T_h를 조정했을 때의 추정율을 나타낸다. 괄호안의 숫자는 식 (3)의 프레임수에

표 5. k와 SNR_{seg}과의 관계

Table 5. Relation between k and SNR_{seg}.

음원	k = 3	k = 6	음원	k = 3	k = 6
M1	**4.97dB**	**-1.05dB**	F1	9.16dB	3.14dB
M2	**5.04dB**	**-0.98dB**	F2	9.95dB	3.93dB
M3	4.13dB	-1.89dB	**F3**	**5.11dB**	**-0.91dB**
M4	3.83dB	-2.19dB	**F4**	**5.85dB**	**-0.17dB**
M5	9.91dB	3.89dB	F5	8.51dB	2.49dB
M6	5.78dB	-0.24dB	F6	9.01dB	2.99dB

대한 비율을 나타낸다. 문턱값이 증가함에 따라서 추정에 사용되어지는 프레임의 수가 증가하는 모양을 알 수 있다. 또한, 다른 문턱값에 대해서도 평균 91% 이상의 높은 추정율이 구해졌다. 그림 5는 샘플수가 각각 128과 256일 경우의 k=0, 3, 6에 대한 추정율의 평균값을 나타내고 있으며, 문턱값 $T_h = r_m/3$의 경우에 대해서 추정율이 가장 높다는 것을 나타내고 있다. 여기에서, N=128, Cep=10에 대한 평균값은 표 6과 동일한 방법으로 구한 잡음율의 추정율이다. 표 4에서 알 수 있듯이, N=128의 경우보다 N=256의 경우가 보다 거리가 떨어져 있으므로, 본 실험에 있어서는 N=256의 경우가 NN에 의한 추정율이 높다는 것을 확인하였다. 따라서 데이터간의 거리의 크기와 NN에 의한 추정율의 사이에는 상관관계가 있는 것이 증명되었다고 할 수 있다.

표 6. T_h에 의한 NN의 추정율(%) (N=256, Cep=20의 경우)

Table 6. Estimation rates of NN based on the difference of T_h's(%) (In the case of N=256, Cep=20).

A: $T_h = r_m/5$의 경우(A: In the case of $T_h = r_m/5$)

추정문장	k=0	k=3	k=6
M1	100(110/110)	99.1(109/110)	100(110/110)
M2	100(75/75)	98.7(74/75)	100(75/75)
F3	98.9(93/94)	97.9(92/94)	100(94/94)
F4	100(80/80)	98.8(79/80)	100(80/80)

B: $T_h = r_m/3$의 경우(A: In the case of $T_h = r_m/3$)

추정문장	k=0	k=3	k=6
M1	100(127/127)	100(127/127)	100(127/127)
M2	100(81/81)	100(81/81)	100(81/81)
F3	99.0(102/103)	100(103/103)	100(103/103)
F4	100(90/90)	100(90/90)	100(90/90)

C: $T_h = r_m$의 경우(A: In the case of $T_h = r_m$)

추정문장	k=0	k=3	k=6
M1	100(178/178)	98.3(175/178)	97.8(174/178)
M2	100(121/121)	96.7(117/121)	97.5(118/121)
F3	97.1(136/140)	91.4(127/140)	97.9(137/140)
F4	96.9(127/131)	90.1(118/131)	100(131/131)

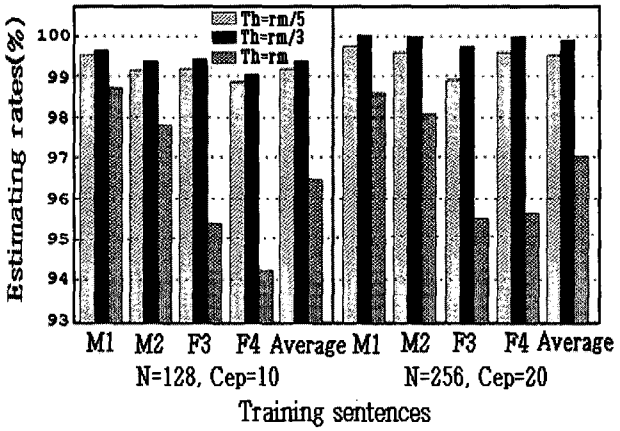

그림 5. 문턱값에 의한 잡음 추정율의 비교

Fig. 5. Comparison of noise estimation rates based on the difference of thresholds.

3. 화자와 문장에 대한 잡음추정 성능의 효과

가. 학습과 추정에 사용한 문장이 동일하고 화자가 다른 경우

표 7은 이 경우의 추정율을 나타낸다. 화자가 다른 경우에도 거의 동일한 정도의 추정율이 구해진다. 이 경우에도 N=256의 경우가 추정율이 높고, 거리와 추정율과의 관계를 증명하고 있다.

나. 화자와 문장이 학습 데이터와 다른 경우

표 8은 4종류의 문장 M1, M2, F3, F4을 사용하여 NN을 학습시킨 후에 M5, M6, F5, F6에 대한 추정율을 나타내고 있고, 추정율의 평균값은 95.8%이다. 이 경우에도, 문턱값 $T_h = r_m/3$의 경우가 다른 문턱값의 경우와 비교해서 가장 양호했다. 또한, 표 6.B의 평균값은 99.9%, 표 7.B의 평균값은 98.0%, 표 8의 평균값은 95.8%의 추정율을 나타낸다. 어느 경우에도 평균 95% 이상의 높은 추정율이 구해져, NN에 의한 잡음추정의

표 7. 화자와 문장이 다른 경우의 NN의 추정율(%)

Table 7. Estimation rates of NN based on the difference of speakers and sentences(%).

A: N=128, Cep=10, $T_h = r_m/3$의 경우

(In the case of N=128, Cep=10, $T_h = r_m/3$)

학습문장	추정문장	k = 0	k = 3	k = 6
M1	F1	95.2%	94.3%	99.5%
M2	F2	89.7%	91.9%	98.5%
F3	M3	95.7%	95.7%	99.4%
F4	M4	97.5%	98.8%	99.4%

B: N=256, Cep=20, $T_h = r_m/3$의 경우

(In the case of N=256, Cep=20, $T_h = r_m/3$)

학습문장	추정문장	k = 0	k = 3	k = 6
M1	F1	96.9%	97.9%	100%
M2	F2	95.5%	97.0%	98.5%
F3	M3	96.2%	96.2%	100%
F4	M4	98.7%	98.7%	100%

표 8. 화자와 문장의 모두가 다른 경우의 NN의 추정율(%)
Table 8. Estimation rates of NN based on the difference of speakers and sentences(%).
N=256, Cep=20, $T_h = r_m/3$의 경우
(In the case of N=256, Cep=20, $T_h = r_m/3$).

학습문장	추정문장	k = 0	k = 3	k = 6
M1	M5	90.9%	96.4%	98.2%
M2	M6	88.7%	96.5%	100%
F3	F7	98.3%	94.8%	96.6%
F4	F8	96.6%	95.6%	96.6%

유효성이 증명되었다.

V. 응용 및 검토

표 9는 k=0~7의 백색잡음이 부가된 경우에 추정문장 M6, F5에 대하여 k=0, k=3, k=6으로 추정되어지는 비율을 나타낸다. 표에 나타난 것처럼, k=0, 1은 k=0으로, k=2, 3, 4는 k=3으로, k=5, 6, 7은 k=6으로 가장 높은 비율로 추정되어지는 것을 알 수 있다. 이 결과를 이용해서 음성강조시스템(참고문헌[5]의 그림 1)의 진폭조정성분계수 R과 상호억제계수 B_f 의 매개변수를 각 프레임의 잡음강도에 따라서 음성을 강조를 하는 것이 본 연구의 최종목표이다. 실제로 이 결과를 음성강조시스템에 적용하면 추정문장 M6에 대하여 잡음량이 k=2의 경우에는 각 프레임에서의 SNR_{seg}이 다르기 때문에 프레임의 33.9%는 k=0으로, 63.5%는 k=3으로, 2.6%는 k=6으로 추정되어 각각의 추정율에 따라서 프레임마다 최적인 R과 B_f의 매개변수가 조정되어 음성을 강조한다. 표 10은 문장 M1, F3에는 백색잡음을, M2, F4에는 교통량이 많은 도로에서 녹음한 자동차의 주행잡음을 부가해 Ⅲ.1절과 동일한 실험 조건하에서 NN을 학습시켰으며, 추정문장 M6에 대해서 k=0~7까지의 백색잡음 또는 자동차의 주행잡음을 부가한 경우에 k=0, k=3, k=6로 추정되어지는 비율을 나타낸다. 여기에서 자동차의 주행잡음의 스펙트럴의 특성은 음성의 스펙트럴의 경사특성과 거의 동일한 특징을 가지는 주파수 분포를 하고 있다. 또한, 자동차의 주행잡음과 백색잡음 모두 크기를 실효값으로 나타내고 있으며, 두 잡음 모두 k=6에 대해서 SNR_{seg}이 거의 0dB이 되도록 실효값을 실험적으로 결정하였다. 표 10의 결과는, 음성에 백색잡음만이 부가된 경우 또는 자동차의 주행잡음이 부가된 경우에, 두 잡음 모두 높은 잡음량의 추정율이 구해지는 것을 나타내고 있다. 이는 자동차의 주행잡음과 같은 음성신호의 스펙트럴에 비교적 유사한 잡음량에 대해서도 본 실험에서의 NN에 의한 잡음량의 추정법이 유효하다는 것을 말할 수 있다. NN을 사용하는 방법 이외에도, 백색잡음과 같은 음성성분과 비교하여 고역성분이 강한 스펙트럴의 대역을 선택적으로 이용하는 등의 잡음량의 추정법을 고려할 수 있다. 그러나 본 연구의 샘플링 주파수는 8kHz이므로 유효한 주파수 대역은 4kHz까지 제한되어 스펙트럴 대역의 선택적인 이용의 효과는 본 연구에서는 기대할 수 없을 것으로 고려된다.

표 9. 잡음량에 대한 NN의 추정율(%)
Table 9. Estimation rates of NN for k=0~7.
N=256, Cep=20, $T_h = r_m/3$의 경우
(In the case of N=256, Cep=20, $T_h = r_m/3$)

학습 문장	백색 잡음량	추정문장 M6			추정문장 F5		
		k=0	k=3	k=6	k=0	k=3	k=6
M1	k=0	**88.7**	10.4	0.9	**98.3**	1.7	0.0
	k=1	**86.1**	13.0	0.9	**96.6**	1.7	1.7
M2	k=2	33.9	**63.5**	2.6	44.8	**53.5**	1.7
	k=3	0.9	**96.5**	2.6	3.5	**94.8**	1.7
F3	k=4	1.7	**84.5**	13.8	0.0	**87.9**	12.1
	k=5	0.0	14.8	**85.2**	0.0	13.8	**86.2**
F4	k=6	0.0	0.0	**100**	0.0	3.4	**96.6**
	k=7	0.0	0.0	**100**	0.0	0.0	**100**

표 10. 잡음량에 대한 NN의 추정율(%)
Table 10. Estimation rates of NN for k=0~7.
N=256, Cep=20, $T_h = r_m/3$의 경우
(In the case of N=256, Cep=20, $T_h = r_m/3$)

학습 문장	백색 잡음량	추정문장 M6			도로 잡음량	추정문장 M6		
		k=0	k=3	k=6		k=0	k=3	k=6
M1	k=0	**87.8**	9.6	2.6	k=0	**87.8**	9.6	2.6
	k=1	**87.0**	10.4	2.6	k=1	**80.9**	13.9	5.2
M2	k=2	42.6	**54.8**	2.6	k=2	31.3	**60.0**	8.7
	k=3	0.9	**94.8**	4.3	k=3	1.7	**95.7**	2.6
F3	k=4	0.9	**79.1**	20.0	k=4	0.0	**62.6**	37.4
	k=5	0.9	14.8	**84.3**	k=5	0.0	14.8	**85.2**
F4	k=6	0.0	0.0	**100**	k=6	0.0	0.0	**100**
	k=7	0.0	0.0	**100**	k=7	0.0	0.0	**100**

VI. 결 론

본 논문에서는 다양한 종류의 잡음의 크기를 추정할

수 있는 신경회로망에 의한 잡음량 추정방식의 시스템을 제안하였다. 제안된 시스템은 화자와 음성 데이터가 학습 데이터와 다르더라도 백색잡음에 대해서 평균 95% 이상의 높은 잡음 추정율을 구할 수 있었다. 실험결과를 정리하면, 다음과 같은 결론을 얻는다.

(1) NN에 의한 잡음량의 추정이 -2dB정도까지의 음성에 대하여 양호하게 학습 가능하다.

(2) 중간층의 유닛수에 따라 NN의 수속성과 추정율이 변한다. 중간층의 유닛수는 30으로 한 경우가 가장 양호하다.

(3) 각 문턱값에 대하여 다른 추정결과가 얻어진다. 문턱값 $T_h = r_m/3$의 경우가 가장 양호하고, $T_h = r_m$ 또는 $T_h = r_m/5$에서도 양호한 추정율이 구해진다.

(4) 화자와 문장이 학습 데이터와 달라도 평균 95% 이상의 높은 추정율이 구해진다.

(5) 본 시스템에 있어서는 샘플수 N=256, 켑스트럼수 Cep=20의 경우가 추정율이 양호하다.

(6) 데이터간의 거리의 크기와 추정율과는 서로 상관관계가 있다.

향후의 연구과제로서는 유색잡음에 대해서도 본 시스템의 유효성을 확인하는 것과 본 시스템의 추정결과를 음성강조시스템에 응용하는 것을 검토할 필요가 있다.

감사의 글

본 연구에 도움을 주신 오사카시립대학 호소카와 교수 및 오카모토 교수에게 감사드립니다.

부 록

음성의 데이터베이스로 사용한 일본어 문장은 다음과 같다. 괄호 안의 문장 들은 일본어를 한국어로 번역한 것이다.

1. M1: "Tsukubawa imakara nijuunenkuraimaeni kenkyuugakuentoshi kousouniyotte tsukurareta hijouni atarashii machidesu"("쯔꾸바는 지금부터 약 20년 전에 연구학원도시 구상에 의해서 만들어진 꽤 새로운 마을입니다.")

2. M2: "Toukyoukara ikouto omoundesuga donoyouna koutsuukikano riyousureba iidesuka"("동경으로부터 출발할려고 합니다만, 어떤 교통기관을 이용하면 좋습니까?")

3. M3: "Kondono onseikenkyuukaio kikini ikitaindesukeredomo douittara iindeshouka"("이번 음성연구회에 참가하러 가고 싶습니다만, 어떻게 가면 좋습니까?")

4. M4: "Kikaishinkoukaikanto iurashiindesuga bashoo zenzen shiranaindesu"("기계진흥회관이라 합니다만, 장소를 전혀 모릅니다.")

5. M5: "Soreninotteitadaite hibiyasenno kamiyachoude orirunoga ichiban chikaito omoimasu"("그것에 승차하셔서 히비야선의 카미야쪼에서 내리는게 가장 가깝다고 생각합니다.")

6. M6: "Kougyougijutsuinno tsukubakenkyuusenta kouenerugi butsurigakukenkyuujo tsukubauchuusenta nadodesu"("공업기술원의 쯔꾸바 연구센터, 고에너지 물리학 연구소, 쯔꾸바 우주센터 등입니다.")

7. F1: M1과 동일한 문장. 8. F2: M2와 동일한 문장.

9. F3: M3와 동일한 문장. 10. F4: M4와 동일한 문장.

11. F5: "Toukyouekide shinkansen oritekara chikatetsuni nottekudasai"("동경역에서 신간센을 내려 가지고 지하철을 타십시오.")

12. F6: "Tsukubano toshimokeio oita kan kouannaijoya taiyounetsuo riyoushita ookina onsuipuruga arimasu"("쯔꾸바의 도시를 모형으로 한 관광 안내소 및 태양열을 이용한 커다란 온수 풀장이 있습니다.")

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. 6, no. 5, pp. 471-472, 1978.

[2] J. S. Lim, A. V. Oppenheim, 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] 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.

[4] Y. M. Cheng, D. O'Shaughnessy, "Speech enhancement based conceptually on auditory

evidence", IEEE Trans. Signal Processing. vol. 39, no.9, pp. 1943-1954, 1991.
[5] 최재승, "청각기강의 모델을 이용한 음성강조 시스템", 전자공학회 논문지 제41권 SP편 제6호, pp. 295-302, 2004.
[6] T. V. Sreenivas, P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement", IEEE Trans. Speech and Audio Processing. vol. 4, no. 5, pp. 383-389, 1996.
[7] S. Oh, V. Viswanathan, P. Papamichalis, "Hands-free voice communication in an automobile with a microphone array", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP vol. 92, no. 1, pp. 281-284, 1992.
[8] 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.
[9] S. Tamura, "An analysis of a noise reduction neural network", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP vol. 89, no. 3, pp. 2001-2004, 1989.
[10] M. R. Sambur, "Adaptive noise cancelling for speech signals", IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, no. 5, pp. 419-423, 1978.
[11] B. Widrow, et al., "Adaptive noise cancelling: Principles and applications", Proc. IEEE, vol. 63, no. 12, pp. 1692-1716, 1975.
[12] A. Ishida, H. gobata, "Speech/Non-speech Discrimination under Real Life Environments". J. Acoust. Soc. Japan, vol. 47, no. 12, pp. 911-917, 1991.
[13] 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 : 음성신호처리, 잡음제거, 신경망, 디지털 TV 등>