오차역전파알고리즘을 사용한 이산푸리에변환에 의한 음성강조 시스템

최재승*

*신라대학교 전자공학과

Speech Enhancement System by Discrete Fourier Transform Using Back-propagation Algorithm

Jae Seung Choi*

*Dept. of Electronic Engineering, Silla University

E-mail : jschoi@silla.ac.kr

요 약

본 논문에서는 신경회로망을 사용하여 이산푸리에변환에 의한 진폭성분과 위상성분을 복원하는 음성강조 시스템을 제안한다. 본 시스템은 신경회로망이 잡음이 부가된 음성신호의 이산푸리에변환의 진폭성분과 위상성분을 사용하여 학습된 후, 제안한 시스템은 배경잡음에 의하여 열화된 잡음이 부가된 음성신호를 강조한다. 배경잡음에 의하여 열화된 음성신호는 신경회로망을 사용하여 제안된 시스템에 의하여 강조되는 것을 실험결과로 증명하며, 제안한 시스템이 스펙트럼 왜곡율의 평가법을 사용하여 배경잡음에 의하여 열화된 음성신호에 대하여 효과적인 것을 실험으로 확인한다.

Ⅰ. 서 론

신경회로망이 실시하는 정보처리는 네트워크 메커니즘에 의한 병렬처리 분산형의 정보처리라고 할 수 있다. 이러한 병렬처리 분산형은 많은 단순한 정보처리요소가 조합되어서 간단한 신호를 주고받고 하는 것과 같은 형태의 네트워크 상의 메커니즘을 사용한 정보처리라고 할 수 있다. 최근에 이러한 네트워크 메커니즘에 의한 정보처리가 인간의 뇌의 구조를 컴퓨터 상에서 모의하는 신경회로망에 의한 정보처리가 여러 방면에서 연구가 활발히 시도되고 있다. 이러한 시도는 뇌와 같은 높은 병렬성을 가진 시스템에의 흥미, 그리고 자기 학습되어 가는 시스템에의 흥미 등에 지지되어 하나의 주류로서 연구의 흐름을 만들어 왔다. 이러한 연구 중에서 도출되어진 것으로는 오차역전파학습법[1, 2]에 의한 신경회로망(Neural Network; NN)[3, 4]이 있다. 본 논문은 이러한 NN을 사용하여 음성신호에 잡음이 중첩된 신호로부터 음성신호의 부분을 강조하는 것을 목적으로 하는 연구이며, 오차역전파학습법에 의한 신경회로망을 사용하여 이산푸리에변환(Discrete Fourier Transform; DFT)에 의한 진폭성분과 위상성분을 복원하는 음성강조 시스템을 제안한다. 먼저, 제안한 신경회로망은 잡음이 부가된 음성신호의 이산푸리에변환의 진폭성분과 위상성분을 사용하여 신경회로망을 학습한 후에 배경잡음에 의하여 열화된 잡음이 부가된 음성신호를 강조한

다. 제안한 시스템은 스펙트럼 왜곡율의 평가법을 사용하여 배경잡음에 의하여 열화된 음성신호에 대하여 효과적인 것을 실험으로 확인한다.

II. 실험에 사용한 네트워크의 구조

본 실험에 사용한 신경회로망은 그림 1에 나타내는 NN이며, Rumelhert [5]에 의해 제안되었던 2승오차 최소화의 학습을 다층의 네트워크 전체의 학습에 확장한 오차역전파 알고리즘의 학습법[1, 2]을 사용한다. 본 실험에서는 입력층이 17유닛, 중간층1이 17유닛, 중간층2가 17유닛, 출력층이 17유닛을 가지는 구조를 가진다.

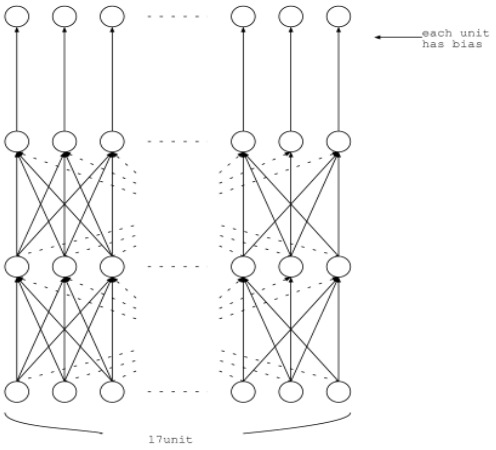

그림 1. 제안한 NN 시스템의 구조

본 실험에서는 32 포인트의 DFT를 실시하므로 실제로 출력되는 계수는 32개이지만, 음성신호 샘플값이 실수 데이터이기 때문에 계수를 다음 식과 같이 하면

$$X(n),\ 0 \le n \le 31 \cdots\cdots\cdots (1)$$

이 계수 사이에는 다음 식의 관계가 성립된다.

$$X(i) = X(32-i)^*,\ 0 \le i \le 15 \cdots\cdots (2)$$

따라서 실제로 입력에 사용되는 DFT 계수는 17개면 충분하다. 또한, 본 네트워크의 학습방법은 오차역전파 학습법을 사용한다. 그리고 이후의 실험에서 NN의 학습횟수는 10,000회로 하여 실험을 실시하였다.

III. 실험 원리 및 NN의 학습방법

본 장에서는 NN의 학습을 이용하여 잡음이 중첩된 DFT 계수의 값을 잡음이 중첩되기 전의 값으로 복원하는 실험에 대하여 기술한다.

본 실험에서 사용한 음성데이터는 일본인 남성 화자에 의한 단어 "aioi(M1)", "hachioji(M2)"를 사용하였으며, 샘플링 주파수는 8 kHz이며 양자화 비트수는 12비트이다. 그리고 백색잡음은 컴퓨터에 의해서 작성한 가우스 잡음을 사용하였으며 샘플링 주파수는 8 kHz이다.

본 실험에 있어서의 평가방법으로서는 스펙트럼 왜곡율(Spectral Distortion; SD)[6]을 사용한다. SD는 입출력 신호의 대수 스펙트럼의 차를 구한 것으로 식 (3)과 같이 정의한다.

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{W}\int_0^w S_x(w) - S_y(w)^2 dw} \cdots (3)$$

단, $S_x(w)$, $S_y(w)$는 입출력 신호의 대수 스펙트럼(dB)이며, N은 측정구간의 프레임 수, W는 신호의 대역폭이다.

본 실험에서는 1프레임을 32 샘플로 하는 음성신호 표본값을 이산푸리에 변환하여 DFT 계수를 구한다. NN에 대한 입력신호로서는 백색잡음을 부가한 음성신호로부터 구한 DFT 계수를, 교사신호로서는 잡음을 부가하지 않은 음성신호로부터 구한 DFT 계수를 각각 부여하여, 오차역전파 학습법에 의하여 NN의 학습을 실시한다. 단, DFT 계수는 실수부와 허수부로 분리하여 각각 별도로 NN을 구축하여 학습을 실시한다. 이 후에, 잡음을 부가한 음성신호로부터 구한 DFT 계수를 학습 종료 후의 NN에 부가하여 출력을 구한다. 이 출력으로서 구해진 DFT 계수를 역이산푸리에변환(Inverse Discrete Fourier Transform; IDFT)하여 음성신호의 샘플값을 구한다. 이 신호와 잡음을 부가하기 전의 음성신호 샘플값를 비교하는 것에 의하여 효과를 조사한다.

본 실험에서는 NN에의 입력인 DFT 값을 0부터 1까지의 범위로 정규화하며, 이를 위하여 본 실험에서 사용한 데이터의 정규화 방법은 다음과 같다. (A) DFT 계수의 실수부 및 허수부의 각 값은 30,000을 부가하여 60,000으로 나눈다. 이렇게 하는 이유는 각 값이 -30,000부터 30,000의 범위에 있다는 가정을 기본으로 하고 있기 때문이다. 실제로 값의 상한은 20,000을 초과하는 것은 없기

때문에 이 방법에 의하여 데이터를 0부터 1의 범위에 제한이 가능하다.

이와 같이 하여 정규화된 각 프레임의 DFT 계수는 입력층의 각 유닛에 부여되기 때문에 다음과 같은 방법으로 학습이 실시된다. (1) 각각의 유닛의 가중치 및 문턱치를 임의로 0부터 1까지의 범위 내에서 초기화하여 학습계수를 지정한다. 본 실험에서는, $\varepsilon = \alpha = 0.3$으로 하였다. (2) 잡음이 중첩된 음성신호의 제1프레임으로부터 구한 DFT 계수를 실수부 및 허수부로 분리하여, 각각 별도의 신경회로망의 입력층의 각 유닛에 부여한다. (3) 잡음이 부가되지 않은 음성신호의 제1프레임으로부터 구한 DFT 계수를 실수부 및 허수부로 분리하여, 각각의 신경회로망에 교사신호로써 부여한다. (4) II장에서 기술한 오차역전파 알고리즘에 의하여 학습을 실시한다. (5) 동일한 방법으로 다음 프레임의 DFT 계수를 입력신호 및 교사신호로써 부여한다. (6) 이하, 동일한 방법의 조작을 반복하여 학습을 실시하며, 모든 프레임에 대하여 학습이 종료된 시점에서 (2)로 되돌아 간다. 단, 되돌아 가는 횟수는 학습횟수로서 지정된 횟수로 한다. 이와 같은 방법으로 학습 후의 네트워크에 잡음이 중첩된 음성신호로부터 구한 DFT 계수를 입력신호로써 부여하면 출력으로써 DFT 계수가 구해지게 된다.

IV. 실험결과

본 장에서는 NN 학습의 일반성을 조사하기 위하여, NN을 사용하여 잡음이 부가된 음성신호를 강조하는 것을 목적으로 한 실험결과에 대해서 기술한다. 표 1은 음성신호를 변경하여도 본 논문에 의한 효과가 구해지는가를 조사하기 위하여, 학습 시에 입력으로 사용한 음성신호와는 다른 음성신호에 백색잡음을 부가한 것을 학습 후의 네트워크에 입력으로써 부여하여 출력을 구하여 효과를 조사하는 실험을 실시하였다. 표 1의 결과로부터 알 수 있듯이 다른 음성에 대해서도 DFT 계수를 NN을 사용하여 잡음이 부가되기 전의 상태로 근접시키는 것에 의하여 SD 경감의 효과가 보이는 것을 알 수 있다.

표 1. 학습 시의 입력신호와는 다른 음성신호를 사용한 경우의 실험결과

Training		After training			
Input speech	Input SD(dB)	Input speech	SD(dB)		
			Input	Output	Impr.
M1	10.47	M2	11.94	10.36	1.58
		M2	14.50	12.24	2.26
		M2	16.18	13.74	2.44
		M3	13.71	10.94	2.77
		M3	16.52	13.31	3.21
		M3	18.57	15.42	3.15
M2	11.94	M1	10.47	8.22	2.25
		M1	12.56	9.16	3.40
		M1	14.68	10.45	4.23
		M3	13.71	8.89	4.82
		M3	16.52	10.81	5.71
		M3	18.57	12.54	6.03

NN의 학습 후에 본 방식에 의한 잡음제거가 효과적인 것을 나타내기 위하여 그림 2, 3, 4를 그래프로 나타낸다. 그림 2는 음성신호 "aioi"의 제3프레임에 대한 NN의 학습신호의 DFT 진폭성분을, 그림 3은 잡음이 부가된 SD=10.47의 음성신호 "aioi"의 제3프레임에 대한 학습하지 않은 시점에서의 NN의 입력신호의 DFT 진폭성분을, 그림 4는 잡음이 부가된 SD=10.47의 음성신호 "aioi"의 제3프레임에 대한 학습한 시점에서의 NN 학습 후의 출력신호의 DFT 진폭성분을 각각 나타낸다. 그림들을 비교함으로서 "aioi"를 학습함으로서 DFT 진폭성분이 충분히 잡음이 없는 상태에 접근한 것을 알 수 있다.

이상의 결과로부터 입력 SD가 최대 18(dB) 정도까지의 잡음에 대해서도 본 방식에 의한 잡음제거 효과가 높은 것을 확인할 수 있었다.

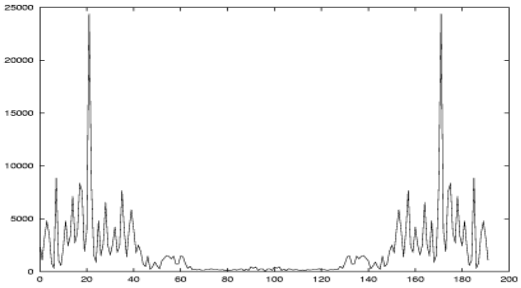

그림 2. NN의 학습신호의 DFT 진폭성분

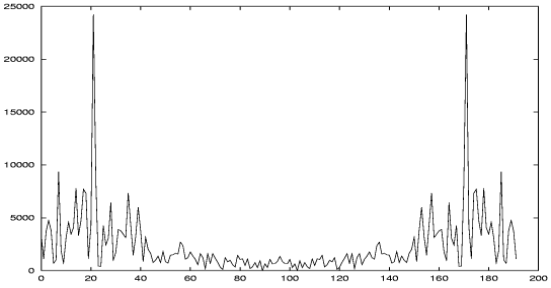

그림 3. NN의 입력신호의 DFT 진폭성분

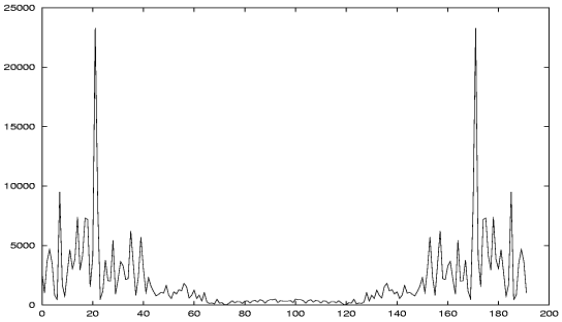

그림 4. 학습 후의 출력신호의 DFT 진폭성분

V. 결론

본 논문에서는 음성강화의 한 방법으로써 오차역전파학습법에 의한 신경회로망을 사용하여 이산푸리에변환에 의한 진폭성분과 위상성분을 복원하는 음성강조 시스템을 제안하였다. 제안한 시스템은 신경회로망의 학습에 의하여 잡음이 부가된 음성신호의 진폭성분과 위상성분을 잡음이 부가되기 전의 음성신호로 복원 가능한 것을 스펙트럼 왜곡율의 평가법을 사용하여 효과적인 것을 실험으로 확인하였다.

향후의 연구과제로는 (1) 신경회로망은 실수부 및 허부수를 별도로 학습하고 있기 때문에 학습에 걸리는 시간을 단축할 필요가 있다. (2) 실험에 사용하는 음성데이터를 다양하게 할 필요가 있다. 이상으로 신경회로망에 의한 DFT 계수의 복원에 의한 본 논문의 수법이 스펙트럴 왜곡율의 값에 의한 개선이 음성신호처리의 분야에 많은 도움이 될 것으로 생각한다.

참고문헌

[1] Ooyen A. V. and Nienhuis B. "Improving the convergence of the back-propagation algorithm," Neural Networks 5, 3, pp. 465-471, 1992.

[2] D. Rumelhart, G. Hinton and R. Williams, "Learning representations by back-propagation errors," Nature323, pp. 533-536, 1986.

[3] S. Tamura, M. Nakamura, “Improvements to the noise reduction neural network”, 1990 International Conference on Acoustics, Speech, and Signal Processing, pp. 825-828, 1990.

[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] D. Rumelhart, "Parallel Distributed Processing, vol. 1 and 2, MIT Press, Cambridge, MA, 1986.

[6] K. Itoh, et al., "A Study of Objective Quality Measures for Digital Speech Waveform Coding Systems", The Institute of Electronics, Information and Communication Engineers(IEICE), Vol. J 66-A, No. 3, pp. 274-281, 1983.