다층 퍼셉트론 신경회로망을 사용한 구간 검출 알고리즘

최재승*

*신라대학교 전자공학과

Section Detection Algorithm using Multi-layer Perceptron Neural Network

Jae Seung Choi*

*Dept. of Electronic Engineering, Silla University

E-mail : jschoi@silla.ac.kr

요 약

본 논문에서는 다층 퍼셉트론 신경회로망을 사용하여 각 프레임에서 유성음, 무성음, 그리고 묵음 구간을 검출하는 구간검출 알고리즘을 제안한다. 신경회로망의 입력으로는 고속 푸리에변환에 의한 전력스펙트럼 및 고속 푸리에변환 계수가 사용되어 네트워크가 학습된다. 본 실험에서는 원 음성에 백색잡음이 중첩된 음성을 신경회로망에 입력함으로서 각 프레임에서의 유성음, 무성음, 묵음 구간의 검출성능 결과를 나타낸다.

키워드

퍼셉트론 신경회로망, 검출 알고리즘, 고속 푸리에 변환

Ⅰ. 서 론

음성은 잡음에 의하여 비선형적으로 변화한다. 이 변화로부터 원래의 음성으로 복구가 가능하다면 음성인식의 전처리로써 사용한다든가 잡음을 제거하는 것이 가능할거라고 생각된다. 잡음이 중첩된 음성파형으로부터 직접 신경회로망 및 선형필터에 의하여 잡음을 제거하는 실험이 보고되고 있다[1, 2]. 그러나 이 수법은 아직 확립되어 있지 않다. 또한 잡음이 중첩된 환경 하에서 음성인식 방식으로서는 신경회로망에 의한 방법(Neural Network, NN)[3, 4, 5, 6], 은닉 마르코프 모델(Hidden Markov Model, HMM)[7, 8]등의 방법들이 연구되고 있다.

본 논문에서는 먼저 신경회로망의 학습 알고리즘을 사용하여 각 프레임에서 유성음, 무성음, 묵음 구간의 검출에 대한 알고리즘을 제안한다. 본 실험에서는 신경회로망에 대해서 입력 신호대잡음비 SNRinput(Input Signal-to-Noise Ratio)을 Clean, 20 dB로 변경한 잡음이 중첩된 음성을 신경회로망에 입력함으로서 각 프레임에서의 유성음, 무성음, 묵음 구간의 검출 결과를 나타낸다.

II. 다층 퍼셉트론 신경회로망

본 논문에서 사용한 신경회로망은 중간층이 1층인 그림 1과 같은 다층 퍼셉트론(Perceptron)[9]형의 계층형 네트워크를 사용하며, 네트워크의 유닛간은 입력층으로부터 출력층으로 향하는 결합을 가진다. 퍼셉트론형의 네트워크에서는 오차역전파 학습 알고리즘[10]을 사용하여 네트워크를 학습시키며, 이 오차역전파 학습 알고리즘의 특징은 교사신호가 있는 학습에 있어서 출력층으로부터 입력층에 오차를 역전파시킴으로써 각 유닛에 대하여 최급강하법을 적용하며, 각 유닛에 비선형함수

를 도입하여 입력으로부터 출력에의 사상을 가능하게 하는 알고리즘이다.

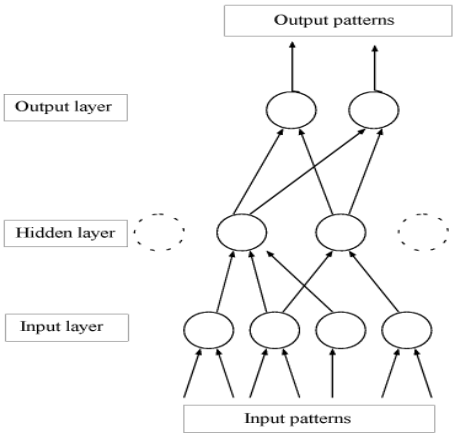

그림 1. 다층 퍼셉트론 신경회로망

III. 데이터베이스 및 평가방법

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어 숫자로 구성된 Aurora2 데이터베이스(Database, DB)를 사용하였다. Aurora2 DB는 남성화자 55명 및 여성화자 55명에 의해서 발성된 음성을 녹음한 총 8440개의 숫자로 구성된 테스트 셋 A, B, C의 음성데이터를 사용하였다. 본 실험에서는 Aurora2 DB 중에서 임의적으로 30문장을 선택하였으며, 10문장은 신경회로망의 학습 데이터로 사용하며 나머지 데이터는 평가용으로 사용하였다.

본 실험에서 사용한 잡음데이터는 컴퓨터에 의해서 작성된 가우스 백색잡음(white noise)의 배경잡음을 사용하여 평가하였다. 이러한 배경잡음은 샘플링 주파수 8 kHz로 (Analog-to-Digital, A/D) 변환한 것을 사용한다. Aurora2 데이터베이스의 각 테스트 셋에 포함된 배경잡음들이 음성데이터에 부가된 잡음이 중첩된 음성데이터(즉, SNR_{input}(Input Signal-to-Noise Ratio) = 20 dB, 15 dB, 10 dB, 5 dB, 0 dB)들이 포함되어 있다. 백색잡음에 대해서는 Aurora2 데이터베이스의 음성데이터에 별도로 백색잡음을 부가하여 잡음이 중첩된 음성데이터를 구하였다.

잡음부가음성은 위에서 기술한 음성데이터에 백색잡음을 중첩하여 작성하였다. 각 음성데이터마다 서로 다른 백색잡음을 중첩함으로써 입력 신호대잡음비(SNR_{input})가 20 dB의 잡음이 중첩된 음성을 작성하였다. 또한, SNR_{input}으로서는 다음 식에서 나타내는 바와 같이 음성 S(n)과 잡음 N(n)의 전체에 해당하는 전력의 비율로서 정의되는 전역 SNR_{input}을 사용하였다.

$$SNR_{input} = 10 \cdot \log_{10}\left(\frac{\sum_{n=1}^{N} S(n)^2}{\sum_{n=1}^{N} N(n)^2}\right) \quad \cdots\cdots\cdots\cdots (1)$$

여기에서, N은 음성데이터의 샘플수이다.

IV. 구간 검출시스템

본장에서는 3층 구조의 퍼셉트론형의 신경회로망에 FFT에 의한 전력스펙트럼 및 FFT에 의한 켑스트럼을 입력으로 하여 각 프레임에서 유성음, 무성음, 묵음에 대한 구간을 검출하는 것을 목적으로 하여 인식율을 높이는 실험에 대하여 기술한다.

그림 2는 본 논문에서 제안하는 학습용 및 평가용의 문장이 동일한 경우의 인식 시스템을 나타낸다. 본 실험에서는 샘플링 주파수 8 kHz의 잡음이 중첩된 이산시간의 음성신호를 128샘플(16 ms)의 프레임으로 분리한 후에 해밍창(Hamming window)을 통과시킨다. Thresholding 블록에서, 해밍창을 통과한 잡음이 중첩된 음성신호는 각 프레임의 실효값 R_f가 문턱값 $R_m/3$보다 큰 경우에는 유성부(모음)으로 판별하도록 하며(즉, $R_f > R_m/3$인 경우), $R_m/5 \le R_f \le R_m/3$일 때에는 이 프레임은 무성부(자음)로 판별하며, $R_f < R_m/5$일 때에는 이 프레임은 묵음부로 각각 판별된다. 여기에서 R_f는 각 프레임에서 구해진 실효값을 나타낸다. 본 실험에서는 처음의 약 5프레임에서 각 문장의 평균 실효값 R_m을 실험적으로 구하였다. 유성부, 무성부, 묵음부로 각각 판별이 된 후에, 각 프레임의 음성신호 표본값으로부터 저역에 해당하는 10차의 FFT 켑스트럼을 구한다. 또한 입력 음성신호에 대해서 FFT를 실시하여 FFT 전력스펙트럼을 구한다. 여기에서 구해진 유성부, 무성부, 묵음부의 데이터에 대하여 정규화하여, 이 데이터들이 3층 구조의 신경회로망의 입력 데이터로 각각 부여되어 유성부, 무성부, 그

리고 묵음부로 판별되도록 신경회로망이 학습된다. 또한 신경회로망의 학습법으로는 오차역전파 학습 알고리즘을 사용한다.

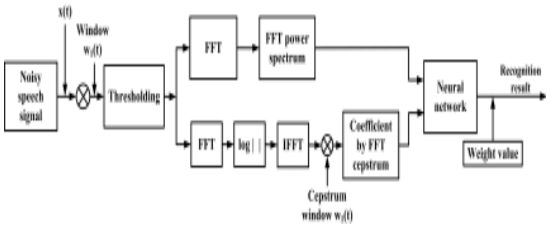

그림 2. 제안한 검출 시스템

제안한 신경회로망 시스템에서는, 입력층의 유닛 수는 10차의 FFT 켑스트럼 계수 및 1개의 FFT 전력스펙트럼의 총 11개를 신경회로망에의 입력으로 사용한다. 신경회로망에의 교사신호는 (O1): 유성부 상태를 [-1.0, 0.0, 0.0], (O2): 무성부 상태를 [0.0, -1.0, 0.0], (O3): 묵음부 상태를 [0.0, 0.0, -1.0]으로 설정하여 유성부, 무성부, 묵음부를 각 프레임에서 인식하도록 신경회로망의 네트워크를 학습시킨다.

따라서 네트워크의 구성은 11개의 입력층 유닛, 15개의 중간층 유닛, 3개의 출력층 유닛으로 구성된 3층의 신경회로망에 입력함으로써, 각 출력신호는 학습신호와 일치한 정확한 값을 취하도록 네트워크를 학습시킨다. 본 실험에서는 학습계수 $\alpha = 0.1$, 가속도 계수 $\beta = 0.6$로 하여 최대 학습횟수를 10,000회로 하였다.

V. 검출시스템의 실험결과

본 실험에서 평가용으로 사용하는 음성은 Aurora2 데이터베이스의 테스트 셋 A, B, C로부터 잡음이 중첩된 음성 데이터들이 임의적으로 선택되었으며, 잡음데이터는 학습 시에 사용한 동일한 잡음인 백색잡음이 선택되었다.

본 실험에서는 신경회로망의 학습을 통해 구해진 가중치의 출력 결합계수를 저장한 후, 학습에 사용한 잡음이 중첩된 음성신호 및 학습에 사용하지 않은 잡음이 중첩된 음성신호의 FFT 전력스펙트럼 및 FFT 켑스트럼계수를 각각 신경회로망의 입력으로 사용하여 교사신호 O1, O2, O3의 목표값과 비교하여 각 프레임에서 검출율을 구한다.

표 1, 2, 3은 신경회로망의 학습데이터로서 음성(M1)만을 사용하여 학습을 실시하여, 학습 시에 사용한 동일한 음성신호(M1) 및 학습 시와 다른 음성신호(M2, F1)를 신경회로망의 입력으로 사용한 경우의 구간 검출율에 대한 실험결과이다. 표 1은 $R_f > R_m/3$, 즉 유성부가 신경회로망에 입력된 경우의 구간 검출율을, 표 2는 $R_m/5 \le R_f \le R_m/3$, 즉 무성부가 신경회로망에 입력된 경우의 구간 검출율을, 표 3은 $R_f < R_m/5$, 즉 묵음이 신경회로망에 입력된 경우의 구간 검출율을 각각 나타낸다. 표 1, 2, 3의 결과로부터 학습데이터 및 평가데이터가 다른 경우에 대해서도 구간 검출율은 최대 92% 이상인 것을 알 수 있다. 본 실험에서는 신경회로망의 학습데이터 및 평가데이터로서 남성화자인 M1, M2 및 여성화자인 F1, F2를 사용하였다.

표 1 음성 학습데이터에 대한 검출율(%)(유성부(즉, $R_f > R_m/3$)이 입력된 경우)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
M1	M1	92.1%	7.9%	0.0%
	M2	89.3%	9.9%	0.8%
	F2	87.6%	11.2%	1.2%

표 2 음성 학습데이터에 대한 검출율(%)(무성부(즉, $R_m/5 \le R_f \le R_m/3$)이 입력된 경우)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
M1	M1	9.4%	90.0%	0.6%
	M2	11.6%	87.1%	1.3%
	F2	12.9%	85.2%	1.9%

표 3 음성 학습데이터에 대한 검출율(%)(묵음부(즉, $R_f < R_m/5$)이 입력된 경우)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
M1	M1	0.8%	11.0%	88.2%
	M2	1.7%	12.9%	85.4%
	F2	2.1%	14.8%	83.1%

표 4는 신경회로망의 학습데이터로서 원 음성(F1)에 백색잡음을 중첩시킨 입력 신호대잡음비 SNRinput=20dB에 대하여 잡음이 중첩된 음성신호를 사용하여 학습을 실시하여, 학습 시에 사용한 동일한 음성신호(F1) 및 백색잡음, 그리고 학습 시와 다른 음성신호(F2, M2) 및 백색잡음을 신경회로망의 입력으로 사용한 경우에 대해서, 유성부($R_f > R_m/3$)가 신경회로망에 입력된 경우의 구간 검출율에 대한 실험결과이다. 표 4의 결과로부터 학습데이터 및 평가데이터가 동일한 경우의 인식율은 최대 86% 이상인 것을 알 수 있다. 그리고 본 실험에는 나타내지 않았지만 무성부 및 묵음부에 대한 실험결과도 유성부의 실험결과와 비슷한 경향이 있었다.

표 4 잡음이 중첩된 음성의 학습데이터에 대한 검출율(%)(유성부(즉, $R_f > R_m/3$)이 입력된 경우)

학습 데이터	평가 데이터	구간 검출율(%)		
		유성음	무성음	묵음
F1	F1	86.3%	13.2%	0.5%
	F2	82.5%	16.2%	1.3%
	M2	81.6%	15.9%	2.5%

VI. 결론

본 논문에서는 다층 퍼셉트론 신경회로망을 사용하여 유성부, 무성부, 그리고 묵음부에 대한 각 프레임에서의 구간 검출에 대한 검출 시스템을 제안하였다. 음성신호 및 입력 신호대잡음비 SNR_{input}(Input Signal-to-Noise Ratio)가 20 dB인 경우에 대하여 유성부, 무성부, 그리고 묵음부의 검출이 유효하다는 것을 실험을 통해서 확인하였다.

이상으로 음성신호 및 잡음이 중첩된 음성신호 중에서 중에서 유성부, 무성부, 묵음부을 검출하는 검출 시스템을 다층 퍼셉트론 신경회로망을 사용하여 본 알고리즘이 백색잡음에 대해서 유효하다는 것을 알 수 있었다.

참 고 문 헌

[1] S. Tamura and A. Waibel, "Noise reduction using connectionist models.", International Conference on Acoustics, Speech, and Signal Processing(ICASSP-88), vol. 1, pp. 553-556, 1988.

[2] T. T. Le, J. S. Mason and T. Kitamura, "Characteristics of multi-layer perceptron models in enhancing degraded speech", Proc. ICSLP-94, pp. 1611-1614, 1994.

[3] W. Y. Huang and R. P. Lippmann, "Neural net and traditional classifiers," Conf. Neural Information Processing Systems, 1987 11.

[4] H. Leung and V. Zue, "Some phonetic recognition experiments using artificial neural nets," ICASSP 88, pp. 422-425, 1988.

[5] W. Y. Huang and R. P. Lippmann, "Comparisons between conventional and neural net and traditional classifiers," IEEE 1st Int. Conf. Neural Network, San Diego, 1987. 6.

[6] R. P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, April 1987.

[7] S. Oberle, A. Kaelin, "HMM-based speech enhancement using pitch period information in voiced speech segments," 1997 IEEE International Symposium on Circuits and Systems, Vol. 4, pp. 2645-2647, 1997.

[8] T. Hirahara and H. Iwamida, "Auditory spectrograms in HMM phoneme recognition," Proc. Int. Conf. Spoken Lang. Process., ICSLP-90, pp. 1381-1384, 1990.

[9] S. K. Pal, S. Mitra, “Multilayer perceptron, fuzzy sets, and classification”, IEEE Transaction on Neural Networks, vol. 3, no. 5, pp. 683-697, 1992.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors," Nature, 323, pp. 533-536, 1986.