시간지연신경회로망을 사용한 잡음 중의 음성인식 수법

최재승*
*신라대학교 전자공학과

Speech Recognition Method under Noisy Environments using Time-Delay Neural Network

Jae Seung Choi*
*Dept. of Electronic Engineering, Silla University
E-mail : jschoi@silla.ac.kr

요 약

잡음환경 하의 회화에서 잡음량을 줄이고 신호처리 시스템의 성능을 향상시키기 위해서는 잡음량에 따라서 적응적으로 처리되는 신호처리 시스템이 필요하다. 또한 잡음이 중첩된 음성으로부터 잡음을 제거하기 위해서는 잡음의 크기에 따라서 음성 처리 시스템의 파라미터를 변경하는 것이 양호한 음질의 음성을 재생하는데 바람직하다. 따라서 본 논문에서는 음성 속에 포함되는 잡음량을 인식하는 방법으로 선형예측계수를 구하여 시간지연신경회로망(Time-delay neural network: TDNN)의 입력으로 사용하여 학습시키는 잡음량을 인식하는 방법을 제안한다. 본 잡음량 인식은 다양한 배경잡음에 의하여 열화된 3종류의 음성이 TDNN에 의하여 학습되어진다. 본 실험에서는 Aurora2 데이터베이스를 사용하여 여러 잡음에 대하여 양호한 인식결과를 확인할 수 있었다.

키워드
Time-delay neural network, speech recognition method, linear predictive coefficient, recognition rate

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. 그러나 음성인식 기술을 상업적으로 적용하기 위해서는 여러 가지 기술적인 문제를 해결해야 한다. 이러한 문제들 중에서 가장 중요한 요소는 음성에 부가되는 배경잡음의 영향을 줄이는 일이다[1].

배경잡음 아래에서의 음성인식에의 응용을 고려한 음성강조 및 잡음제거를 위한, 스펙트럼 차감법(spectral subtraction)[2], [3], [4], 위너필터(Wiener filter)[5], microphone array[6], [7], 신경회로망[7], [8], 적응 필터법[9], [10] 등의 방식이 발표되었다. 이러한 논문 중에 신경회로망(Neural Network, NN) 기법은 음성인식에 있어서 상당히 효과적인 능력이 있으며, 음성구간 및 문자의 인식에 대해서도 많은 성과를 올리고 있다[1]. 또한 음성 중에서 잡음을 경감하기 위해서는 잡음의 강도에 따라서 각각 적당한 처리를 할 필요가 있다. 즉, 잡음의 크기를 인식하는 것이 상당히 중요하다.

본 논문에서는 NN에 시간요소를 도입한 시간이 지연된 신경회로망(time-delay neural network: TDNN)[11]을 사용한다. 본 논문에서는, 배경잡음

의 영향을 줄여서 음성인식 시스템의 성능을 향상시키고 다양한 음성인식기의 입력으로 사용하기 위하여, 선형예측분석에 의한 선형예측계수를 시간지연신경회로망의 입력으로 한 시스템을 구축하고자 한다. 본 논문에서 사용하는 시간지연신경회로망의 입력데이터로는 각각의 프레임의 데이터를 사용하여 학습시키며, 시간지연신경회로망의 학습조건 및 학습방법 등을 바꾸어 음성 중의 잡음량을 인식하여 이러한 잡음을 경감하는 것을 목적으로 한 연구를 진행한다. 본 연구의 목적을 달성하기 위하여 본 논문에서는 잡음과 음성신호의 특징을 가진 선형예측계수(Linear Predictive Coefficient, LPC)[12]를 시간지연신경회로망의 입력으로 하여 3종류의 잡음량을 인식하는 방법을 제안한다.

II. 제안한 시간지연 신경회로망의 구조

일반적인 TDNN의 구조는 그림 1과 같다. 입력층의 가로축 방향은 시간을 나타내며, 여기에서 시간 축 방향의 최소단위를 1프레임이라고 한다. 그림은 입력층의 3프레임 분의 유닛(unit)은 중간층 제1층의 1프레임분의 유닛에 결합된다. 중간층 제1층의 4프레임분의 유닛은 중간층 제2층의 1프레임분의 유닛과 결합된다. 중간층 제2층의 각 가로 1열의 유닛은 출력층의 각 유닛에 대응하여 접속되어 있다. 이러한 구조를 가짐으로써 신호의 시간변화 패턴을 반영할 수 있는 네트워크가 구축 가능하다.

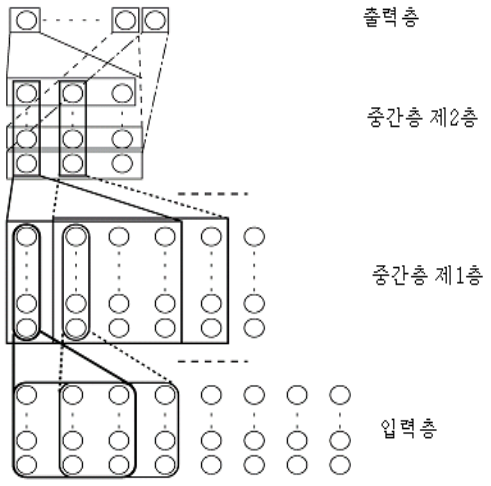

그림 1 제안한 TDNN의 구조

제안한 TDNN은 역전파알고리즘을 사용하여 학습한다. 이 학습법은 입력값이 주어졌을 때 교사신호와 출력값의 오차를 최소로 하며 신경세포 사이의 결합계수를 조절하는 방법이다. 본 실험에서는, 잡음이 없는 음성(SNRin(Input Signal-to-Noise Ratio)=∞), 잡음이 적은 음성(SNRin=15dB), 잡음이 많은 음성(SNRin=5dB)의 3종류를 인식할 수 있도록, 신경회로망의 출력층의 유닛수를 3으로 하여 학습을 시킨다. 이것으로부터 구해진 결과는 실제의 잡음량과 비교되어 인식율의 형태로 평가한다.

이상과 같은 음성에 의하여 구해진 12차의 선형예측계수는 입력층의 각 유닛에 입력되며, 신경회로망의 교사신호는 (T1) SNRin=∞일 때 (1.0, -1.0, -1.0), (T2) SNRin=15dB 일 때 (-1.0, 1.0, -1.0), (T3) SNRin=5dB 일 때, (-1.0, -1.0, 1.0)으로 한다. 그리고 각각의 유닛의 하중은 -0.009~+0.009의 범위 내의 램덤한 값으로 초기화하며, 학습계수 $\alpha = 0.1$, 가속도계수 $\beta = 0.6$로 한다. 본 실험에서는, 4층 구조의 퍼셉트론(perceptron)형[13]의 신경회로망의 구조인 입력층의 12 유닛, 제1 중간층의 20 유닛, 제2 중간층의 20 유닛, 출력층의 3 유닛으로 구성된 네트워크를 사용하였다. 또한 학습의 횟수를 10,000회로 하여 각 음성데이터에 있어서 결합하중의 초기값을 바꾸어서 10회씩 시행한다.

III. 음성신호의 선형예측분석에 의한 잡음량 인식

음성신호의 표본값 사이에는 커다란 상관관계가 있으며 음성신호의 특징을 추출하기 위하여 이것을 이용한 예측부호화가 실시되어진다[12]. 본 논문에서는, 배경잡음의 영향을 줄여서 음성인식 시스템의 성능을 향상시키고 다양한 음성인식기의 입력으로 사용하기 위하여, 선형예측분석에 의한 선형예측계수를 시간지연신경회로망의 입력으로 한 시스템을 구축하고자 한다. 구해진 선형예측계수는 분석의 대상인 일련의 데이터를 전극 모델에 의하여 생성하였을 때의 시스템의 요소가 된다. 따라서 이 계수로서 생성 모델의 정보가 추출되어, 이것들을 부호화함으로서 고능률 부호화가 가능하다.

본 논문에서 제안한 잡음량 인식 시스템을 그림 2에 나타낸다. 표본 주파수 8 kHz의 이산 시간신호 x(t)를 해밍창에 의해 256 표본의 프레임으로 분리한다. 각 프레임의 표본값을 선형예측 분석하여 12차의 선형예측계수를 구한다. 이렇게 함으로써 원래의 표본값은 12차의 선형예측계수와 잔차신호로 완전히 복구가 가능하다. 그리고 12차의 선형예측 계수를 시간지연신경회로망의 입력으로 사용하여 학습을 한다. 시간지연신경회로망의 학습을 통해 얻어진 가중치를 저장한 후, 학습에 사용되지 않은 잡음이 중첩된 음성데이터의 선형예측 계수를 시간지연신경회로망의 입력으로 받아 교사신호 T1, T2, T3의 목표치와 비교하여 각 프레임에서 잡음량을 인식한다.

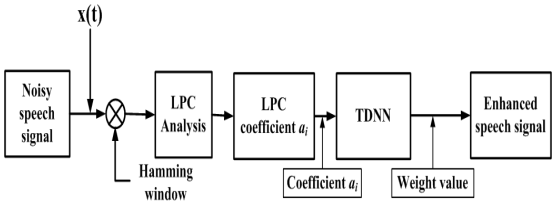

그림 2. 잡음량 인식 시스템

본 실험에서 사용한 음성 데이터는 8 kHz의 표본화 주파수를 가진 환경에서 녹음된 연결된 영어숫자로 구성된 Aurora2 데이터베이스[14]이다. 제안한 시스템은 Aurora2 데이터베이스로부터의 테스트 셋 A, B, C의 음성데이터와 테스트 셋 A의 자동차(car noise), 지하철잡음(subway noise), 그리고 컴퓨터에 의해서 작성된 가우스 백색잡음(white noise) 등의 배경잡음을 사용하여 평가하였다. 본 실험에서는 3종류의 입력 신호대잡음비(SNRin=∞, 15 dB, 5 dB)와 같이 잡음이 부가된 음성신호를 사용하여 시간지연신경회로망을 학습시켰다. Aurora2 데이터베이스를 사용할 경우에 백색잡음, 자동차잡음, 지하철잡음을 Aurora2 데이터베이스의 음성신호에 부가한 후에 시간지연신경회로망이 학습되었다.

Ⅳ. 실험 결과

본 논문에서는 출력되어진 학습결과와 학습신호를 비교하여 인식율을 구한다. 인식율은 전체 프레임 수에 대한 정확하게 인식된 프레임수를 백분율로 나타낸다.

표 1은 Aurora 2 데이터베이스의 테스트 셋 C로부터 임의적으로 20개의 문장을 선택하여, 각 음성으로부터 구한 선형예측계수를 시간지연신경회로망의 입력으로 하여 실험을 실시한 학습 결과를 각 잡음에 대하여 나타낸 평균이다.

표 1. 각 잡음에 대한 잡음량 인식율

Type of noise	Recognition rates (%)		
	T1 (∞)	T2 (15 dB)	T3 (5 dB)
white	99.6%	98.9%	100.0%
Car	99.1%	96.5%	99.3%
Subway	97.9%	96.0%	98.4%
Average	**98.9%**	**97.1%**	**99.2%**

본 논문에서 제안한 선형예측계수에 의한 인식율은 여러 잡음에 대하여 평균적으로 약 98.4% 이상의 높은 인식결과를 확인할 수 있었다. 또한 본 논문에서 제안한 표 1의 선형예측계수에 의한 3 패턴의 학습신호에 의한 학습결과로부터, SNRin=15dB(T2)의 인식율이 다른 입력(T1 및 T3)보다 약간 인식율이 떨어지는 반면에, SNRin=∞(T1)과 SNRin=5dB(T3)에서 상당히 좋은 인식결과를 볼 수 있었다.

Ⅳ. 결론

본 논문에서는 시간지연신경회로망에 의한 3종류의 음성신호의 잡음량을 인식하는 것을 목적으로 하여, 선형예측계수를 입력으로 한 잡음량 인식의 실험을 실시하여 높은 잡음량 인식율의 결과를 확인할 수 있었다. 그러나 향후에는 다양한 실생활에서의 유색잡음 들을 사용하여 실험을 할 필요가 있다고 본다.

이상과 같이 잡음이 중첩된 음성신호에 대한 잡음량의 인식을 시간지연신경회로망을 통하여 실험적으로 확인하여 본 연구가 음성인식 분야에 효과적으로 응용될 것이라고 생각한다.

참고문헌

[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, April 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] 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.
[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, P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement", IEEE Trans. Speech and Audio Processing. vol. 4, no. 5, pp. 383-389, 1996.
[6] 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, vol. 92, no. 1, pp. 281-284, 1992.
[7] 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.
[8] S. Tamura, "An analysis of a noise reduction neural network", IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 89, no. 3, pp. 2001-2004, 1989.
[9] M. R. Sambur, "Adaptive noise cancelling for speech signals", IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, no. 5, pp. 419-423, 1978.
[10] B. Widrow, et al., "Adaptive noise cancelling: Principles and applications", Proc. IEEE, vol. 63, no. 12, pp. 1692-1716, 1975.
[11] M. Miyatake, H. Sawai, and K. Shikano, "Training Methods and Their Effects for Spoting Japanese Phenomes Using Time-Delay Neural Networks", IEICE, Vol. J73-D-II, No.5, pp. 699-706, 1990.
[12] P.B. Patil: Multilayered network for LPC based speech recognition. IEEE Transactions on Consumer Electronics, Vol. 44, No. 2, pp. 435 - 438, 1998.
[13] S. K. Pal, S. Mitra, "Multilayer perceptron, fuzzy sets, and classification", IEEE Transaction on Neural Networks, vol. 3, no. 5, pp. 683-697, 1992.
[14] 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.