FFT 켑스트럼을 사용한 배경잡음의 제거

최재승*
*신라대학교 전자공학과

Reduction of Background Noise using FFT cepstrum

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

요 약

본 논문에서는 오차역전파 학습 알고리즘을 사용하여 신경회로망을 학습시켜, 각 프레임에서의 음성 및 잡음 구간의 검출에 의한 음성인식 알고리즘을 제안한다. 그리고 신경회로망에 의하여 음성 및 잡음 구간의 검출에 따라서 각 프레임에서 잡음을 제거하는 스펙트럼 차감법을 제안한다. 본 실험에서는 원음성에 백색잡음 및 자동차잡음을 부가하여 음성인식의 인식율을 평가한다. 또한 인식시스템에 의하여 검출된 음성 및 잡음 구간을 이용하여 각 프레임에서의 스펙트럼 차감법에 의한 잡음제거의 실험결과를 나타낸다.

키워드
배경잡음, 음성인식 알고리즘, 신경회로망

Ⅰ. 서 론

최근, 신경회로망을 사용한 음성인식을 실시하려고 하는 연구가 활발히 진행되며, 이러한 수법이 음성인식과 같은 일종의 애매함을 포함하는 문제의 해결에 유효함이 해결되어 왔다. 이 중에서도 오차역전파 학습 알고리즘을 사용한 방법은 비교적 간단한 알고리즘임에도 불구하고, 패턴 인식에 있어서 상당히 강력한 학습 알고리즘이라는 것이 다수의 연구에 의하여 증명되고 있다[1, 2, 3, 4]. 그러나 음성인식이 실용화되기 위해서는 아직 해결해야할 여러 가지 문제점이 남겨져 있다. 예를 들면, 마이크로폰에 있어서 주변으로부터 혼입되는 잡음의 영향에 의한 잡음제거의 문제, 불특정 화자화, 음운 및 음절이 연속적으로 발성된 경우에 일어나는 조음결합 현상 등이다. 이 중에서도 잡음의 영향에 의한 잡음제거에 대해서는 각종 실용화의 장면을 고려하면 음성과 잡음의 혼재를 방지하는 것은 거의 불가능하며, 배경잡음을 제거하는 수법이 음성인식의 전처리로써 필요하다.[5 6, 7]

본 논문에서는 먼저 신경회로망의 학습에 오차역전파 학습 알고리즘을 사용하여 각 프레임에서의 음성 및 잡음 구간의 검출에 의한 인식 알고리즘을 제안한다. 그리고 신경회로망에 의하여 음성 및 잡음 구간의 검출에 따라서 각 프레임에서 잡음을 제거하는 스펙트럼 차감법[8]을 제안한다.

본 실험에서는 신경회로망에 대해서 입력 신호 대잡음비 SNRinput(Input Signal-to-Noise Ratio)을 ∞, 10 dB, 5 dB, 0 dB로 변경한 원음성 및 잡음이 중첩된 음성을 신경회로망에서 평가함으로서 인식결과를 나타낸다. 또한 인식시스템에 의하여 검출된 음성 및 잡음 구간을 이용하여 각 프레임에서의 잡음제거의 실험결과를 나타낸다.

II. 제안한 인식 시스템 및 데이터베이스

본 논문에서는 입력층, 중간층 및 출력층으로 구성되는 다층 퍼셉트론[9, 10]형의 계층적인 신경회로망을 제안한다. 퍼셉트론형의 네트워크에서는 오차역전파 학습 알고리즘[11, 12]을 사용하여 네트워크를 학습시키며, 이 오차역전파 학습 알고리즘의 특징은 교사신호가 있는 학습에 있어서 출력층으로부터 입력층에 오차를 역전파시킴으로써 각 유닛에 대하여 최급강하법을 적용하며, 각 유닛에 비선형함수를 도입하여 입력으로부터 출력에의 사상을 가능하게 하는 알고리즘이다.

본 절에서는 3층 구조의 퍼셉트론형의 신경회로망에 고속 푸리에 변환(fast Fourier transform : FFT)에 의한 전력스펙트럼 및 FFT에 의한 켑스트럼을 입력으로 하여 각 프레임에서 음성 및 잡음에 대한 구간을 검출하는 것을 목적으로 하는 인식시스템에 대하여 기술한다.

그림 1은 본 논문에서 제안하는 인식 시스템을 나타낸다. 본 실험에서는 샘플링 주파수 8 kHz의 이산시간신호를 128샘플의 프레임으로 분리하여 각 프레임의 샘플값을 해밍창를 통과시킨 후에 켑스트럼 변환(FFT➔log| |➔IFFT)을 한다. 구해진 켑스트럼을 켑스트럼창에 통과시킴으로써 저역 부분의 10개의 켑스트럼 계수를 구한다. 또한 입력 데이터에 대해서 FFT를 실시하여 FFT 전력 스펙트럼을 구한다. 따라서 제안한 신경회로망 시스템에서는, 입력층의 유닛수는 10개의 켑스트럼 및 1개의 전력 스펙트럼의 총 11개를 신경회로망에의 입력으로 한다. 여기에서 구해진 데이터에 대하여 -0.05～+0.05의 사이에 해당하는 값으로 정규화하여, 이것을 3층 구조의 신경회로망의 입력 데이터로 부여하여 시뮬레이션을 실시한다.

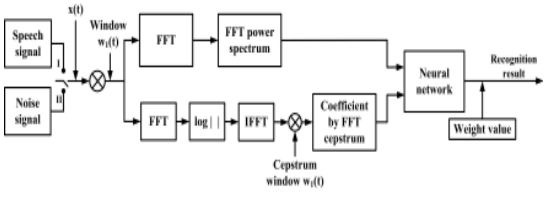

그림 1. 제안한 인식 시스템

제안한 신경회로망에의 교사신호는 (O1): 음성 상태를 [-1.0, 0.0, 0.0], (O2): 백색잡음인 상태를 [0.0, -1.0, 0.0], (O3): 자동차잡음(car noise)인 상태를 [0.0, 0.0, -1.0]으로 설정하여 음성신호, 백색잡음, 그리고 자동차잡음을 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 평균 2승 오차의 변화가 거의 없어지는 10,000회로 하였다.

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어 숫자로 구성된 Aurora2 데이터베이스(Database, DB)[13]의 테스트 셋 A, B, C의 음성데이터를 사용하였다. 본 실험에서는 Aurora2 DB 중에서 임의적으로 20문장을 선택하였으며, 10문장은 신경회로망의 학습 데이터로 사용하며 나머지 데이터는 평가용으로 사용하였다. 또한 본 실험에서 사용한 잡음데이터는 Aurora2 데이터베이스의 테스트 셋 A의 자동차잡음(car noise)을 사용하였으며, 컴퓨터에 의해서 작성된 가우스 백색잡음(white noise)의 배경잡음을 사용하여 평가하였다. Aurora2 데이터베이스의 각 테스트 셋에 포함된 배경잡음들이 음성데이터에 부가된 잡음이 중첩된 음성데이터(즉, SNR_{input} = 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.

III. 배경잡음의 제거 시스템

본 논문에서 제안하는 음성 및 잡음구간 인식 시스템에 의한 스펙트럼 차감법에 의한 신호처리 블록도를 그림 2에 나타낸다.

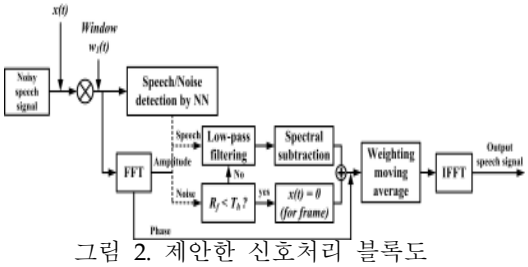

그림 2. 제안한 신호처리 블록도

먼저 잡음이 중첩된 음성신호를 샘플링 주파수 8 kHz로 A/D 변환된 이산시간신호 $x(t)$는 128 샘플을 1 프레임으로 하여 신호처리를 실시한다. 입력 음성신호는 해밍창을 곱한 후에 FFT로 스펙트럼 변환된다. 제안한 인식 시스템에 의하여 각 프레임에서 음성 및 잡음 구간이 검출된다. 각 프레임에서 검출된 음성 및 잡음 구간을 이용하여, 음성 구간인 경우에는 저역통과 필터를 통과시킨 후에 스펙트럼 차감법에 의하여 잡음을 제거한다. 또한 잡음 구간일 때에는 $R_f < T_h$인 경우(즉, 잡음으로 판단)에 한하여 각 프레임에서 입력 신호를 0으로 하며, $R_f \geq T_h$인 경우(즉, 음성으로 판단)에는 음성 구간으로 판단하여 음성으로 재 추정하여 위에서 기술한 바와 같이 저역통과 필터 및 스펙트럼 차감법에 의하여 잡음을 제거한다. 이 후에 음성 및 잡음 구간에 대한 신호를 합성하여 가중치가 부여된 이동 평균 및 IFFT(Inverse Fast Fourier Transform)를 실행함으로써 잡음이 제거된 출력 신호를 구한다.

본 실험에서는 각 문장 전체에서 구한 평균 실효값 R_m을 구하여, 이 실효값의 $R_m/3$값이 문턱값 T_h가 되도록 실험적으로 결정하였다. 즉, 각 프레임에서 $R_f < T_h$일 때에는 이 프레임은 잡음으로 판별되며, $R_f \geq T_h$일 때에는 이 프레임은 음성으로 판별된다. 여기에서 R_f는 각 프레임에서 구해진 실효값을 나타낸다.

본 실험에서는 처음의 약 6프레임을 잡음 구간으로 추정하여 음성신호로부터 이 잡음신호를 차감하는 스펙트럼 차감법을 사용한다.

IV. 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 백색잡음 및 Aurora2 데이터베이스의 테스트 셋 A의 자동차잡음(car noise)이 선택되었다.

본 실험에서는 신경회로망의 학습을 통해 구해진 가중치의 출력 결합계수를 저장한 후, 학습에 사용한 잡음이 중첩되지 않은 원 음성 및 잡음이 중첩된 음성신호의 FFT 전력스펙트럼 및 FFT 켑스트럼 계수를 각각 신경회로망의 입력으로 사용하여 교사신호 O1, O2, O3의 목표값과 비교하여 각 프레임에서 인식율을 구한다.

표 1은 신경회로망의 학습을 통해 구해진 가중치의 출력 결합계수를 사용하여, 다양한 잡음레벨들(SNRinput= Clean, 10 dB, 5 dB, 0 dB)에 대하여 학습에 사용한 동일한 음성신호에 백색잡음을 신경회로망의 입력으로 사용하여 10개의 음성 및 잡음데이터에 대하여 평가 한 경우의 각 프레임에서의 인식율을 나타낸다. 표에서 Clean은 잡음이 중첩되지 않은 원 음성을 의미한다.

표 1 백색잡음인 경우의 학습 데이터와 평가 데이터가 같은 경우의 인식율[%]

SNR_{input} (dB)	인식율[%]
Clean	82.6%
10 dB	63.2%
5 dB	52.7%
0 dB	44.1%
평균	60.7%

표 1의 결과로부터 알 수 있듯이, 잡음이 중첩되지 않은 원 음성의 경우(Clean)에는 양호한 인식율이 구해졌지만, SNRinput이 0 dB에 가까우면, 즉 잡음이 많이 중첩된 음성일 경우에는 인식율이 나쁘게 되어 음성 및 잡음 구간을 검출하는데 어려움이 있는 것을 알 수 있다.

표 2는 다양한 잡음레벨들(SNRinput=10 dB, 5 dB, 0 dB)에 대하여 학습에 사용한 음성신호 및 백색잡음을 입력으로 한 경우에 대하여, 10개의 음성 및 잡음데이터에 대하여 평가한 경우의 SNRoutput에 의한 잡음제거에 대한 실험이다. 표 2의 결과로부터 SNRoutput이 최대 6.8 dB 정도 향상되는 것을 확인할 수 있었다. 표에는 나타내지 않았지만 자동차잡음에 대한 인식율 및 SNRoutput 개선량은 백색잡음에 비하여 평균 2% 및 1 dB 이상 좋지 않게 나타나는 것을 알 수 있었다.

표 2. SNRinput=10 dB～0 dB인 경우의 학습 시의 입력데이터와 학습 후의 입력데이터가 동일한 경우의 결과

SNRinput [dB]	SNRoutput [dB]		
	Classical algorithm of SS	Proposed algorithm	Impr.
10 dB	10.9 dB	16.6 dB	4.7 dB
5 dB	7.8 dB	13.1 dB	5.3 dB
0 dB	6.1 dB	12.9 dB	6.8 dB

V. 결론

본 논문에서는 신경회로망의 오차역전파 학습 알고리즘에 의한 음성 및 잡음 구간의 인식, 그리고 신경회로망에 의하여 음성 및 잡음 구간의 검출에 따라서 각 프레임에서 잡음을 제거하는 스펙트럼 차감법을 제안하였다. 입력 신호대잡음비 SNR_{input}(Input Signal-to-Noise Ratio)이 Clean, 10 dB, 5 dB, 0 dB로 입력된 음성 및 잡음이 중첩된 음성에 대해서 본 알고리즘이 유효하다는 것을 실험에서 확인하였다.

참고문헌

[1] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Comput. Speech Lang. vol. 1, pp. 109-130, 1986.

[2] O. Ghitza, "Auditory neural feedback as a basis for speech processing," IEEE Int. Conf. Acoust. Speech Signal Process., pp. 91-94, 1988.

[3] H. Hamada, T. Hirahara, A. Imamura, T. Matsuoka and R. Nakatsu, "Auditory-based filter-bank analysis as a front-end processor for speech recognition," Proc. Eurospeech 89, Vol. 2, pp. 396-399, 1989.

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

[5] Simpson, et. al., "Spectral Enhancement to Improve the Intelligibility of Speech in Noise for Hearing Impaired Listeners", Acta Otolaryngol, Suppl. 469, pp. 101-107, 1990.

[6] J. P. Haton, "Automatic recognition of noisy speech," in Speech Recognition and Coding New Advances and Trends, A.J.R. Ayuso and J.M.L. Soler, Eds. (Springer-Verlag, Berlin, 1995), pp. 3-13.

[7] H. Hamada, T. Hirahara, A. Imamura, T. Matsuoka and R. Nakatsu, "Auditory-based filter-bank analysis as a front-end processor for speech recognition," Proc. Eurospeech 89, Vol. 2, pp. 396-399, 1989.

[8] 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.

[9] 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.

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

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

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

[13] 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.