주파수대역별 TDNN을 이용한 음성신호의 잡음억제

최재승*

*신라대학교 전자공학과

Noise Suppression of Speech Signal using TDNN for each Frequency Band

Jae Seung Choi*

*Dept. of Electronic Engineering, Silla University

E-mail : jschoi@silla.ac.kr

요 약

본 논문에서는 신경회로망(Neural network)에 시간구조를 도입한 시간지연 신경회로망(Time-delay Neural Network: TDNN)을 사용하여 잡음을 포함한 음성신호로부터 잡음을 제거함으로써 음성을 강조하는 것을 목적으로 한다. 본 논문에서는 먼저 각 프레임의 FFT 진폭성분들을 유성음 구간과 무성음 구간으로 검출한 후, 무성음 구간에 대해서는 각 프레임에서 이동평균을 취하여 음성을 강조한다. 유성음 구간에 대해서는 각 프레임의 FFT 진폭성분들을 저역, 중역 및 고역으로 각각 분리한 후에 각 대역의 FFT 진폭성분들을 저역용 TDNN, 중역용 TDNN, 그리고 고역용 TDNN의 입력으로 하여 각 TDNN에 학습시킴으로써 최종 FFT 진폭성분들을 구한다. 본 실험에서는 Aurora2 데이터베이스를 사용하여 FFT의 진폭성분을 복원하는 잡음제거의 알고리즘을 사용하여 여러 잡음에 대해서 본 알고리즘의 유효성을 실험적으로 확인한다.

Ⅰ. 서 론

음성강조 및 잡음제거의 목적을 위하여 spectral subtraction[1][2], 신경회로망(neural network: NN)[3][4], adaptive noise cancelling[5] 등의 방법 들이 발표되어 다방면에서 연구되고 있다. 이러한 논문 중에 NN이 음성신호처리 분야에 사용되는 경우에의 주요한 응용분야는 음성인식 분야에서 주로 이용되며, 음성강조 및 잡음제거의 목적으로 사용되는 경우에는 잡음이 중첩된 음성구간에서 유성음 및 무성음의 추출[6][7] 등의 전 처리적인 역할을 담당하는 경우와 환경음 인식 등의 경우가 많다. 일반적으로 다층 퍼셉트론 NN에 의한 역전파(back propagation : BP) 알고리즘이 강력한 알고리즘임을 생각한다면 NN이 음성신호처리 분야에의 응용도 유망하다고 고려된다. 따라서 본 논문에서는 NN의 기본 개념을 적용하여 잡음이 중첩된 음성신호의 공간으로부터 잡음이 없는 음성신호의 공간으로 사상을 실행함으로써 잡음을 제거하는 것을 목적으로 한다.

본 논문에서는 NN에 시간요소를 도입한 시간이 지연된 신경회로망(time-delay neural network: TDNN)[8]을 사용한다. 또한 음성신호를 고속 푸리에 변환(fast Fourier transform : FFT)한 경우 위상성분보다 진폭성분이 음성 정보를 많이 포함하고 있다. 따라서 본 논문은 스펙트럼 회복의 수단으로써 TDNN을 이용하여 FFT 진폭성분을 복원하는 알고리즘을 제안한다. 본 실험에서는 Aurora2 데이터베이스를 사용하여 FFT의 진폭성분을 복원하는 잡음제거의 알고리즘을 사용하여 여러 잡음에 대해서 본 알고리즘의 유효성을 실험적으로 확인한다.

II. 제안한 알고리즘

본 실험에서의 음성 및 잡음 데이터베이스는 8 kHz의 표본 주파수를 가지며 영어숫자로 구성된 Aurora2 데이터베이스를 사용한다. 제안한 시스템은 Aurora2 데이터베이스(테스트 셋 A, B, C 포함)로부터의 테스트 셋 A와 B의 음성데이터와 컴퓨터에 의해서 생성한 가우스 백색잡음을 사용하여 평가하였다. 본 실험에서 다양한 신호대잡음비(Signal-to-Noise Ratio: SNR=20dB, 15dB, 10dB, 5dB)이 부가된 잡음이 중첩된 음성신호를 사용하여 TDNN이 학습되었다. 본 실험에서 사용한 배경잡음의 데이터는 샘플링 주파수 8 kHz이다. 또한 본 실험에서의 음성 및 잡음의 주파수 대역은 0 ~ 3kHz이다.

본 논문에서는 주파수 영역에서 스펙트럼 복귀의 한 방법으로써 주파수 대역별 TDNN을 사용하여 FFT 진폭성분을 복귀하는 알고리즘을 제안한다. 본 논문에서는 각 주파수 대역의 스펙트럼 간에 상관이 적은 정보를 TDNN에 제공하여 학습의 효과를 높이기 위하여 저역부, 중역부 및 고역부에 해당하는 개선된 주파수 대역별 TDNN의 알고리즘을 그림 1과 같이 제안한다.

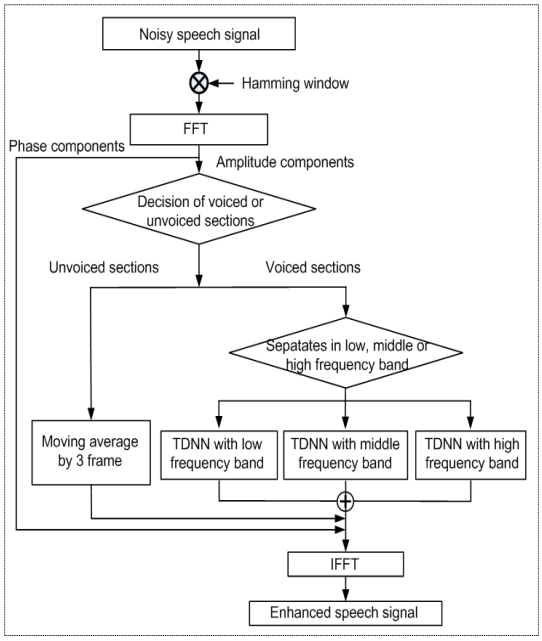

그림 1 제안한 주파수 대역별 TDNN 시스템

먼저 배경잡음이 중첩된 음성신호는 한 프레임이 128 샘플로 구성되며 각 프레임에 대하여 해밍창(hamming window)를 곱하여 FFT에 의하여 진폭성분 및 위상성분들로 분리된다. 진폭성분들은 유성음 및 무성음으로 분리된다. 각 프레임에서, $R_f \geq T_h$일 때에는 이 프레임은 유성음으로, $R_f < T_h$일 때에는 이 프레임은 무성으로 판별된다. 여기에서 R_f는 각 프레임에서 구해진 실효값을 나타낸다. 본 실험에서는 음성구간의 처음의 약 5프레임에서 각 음성데이터의 평균 실효값 R_m을 구하여 이 실효값이 문턱값 T_h가 되도록 실험적으로 결정하였다. 각 프레임에서 유성음 및 무성음으로 판별된 후에, 유성음에서는 FFT의 진폭성분들이 저역, 중역, 고역으로 분리되며 분리된 FFT 진폭성분들은 각 대역의 TDNN의 입력으로 부가되어 그림 1의 TDNN에 의하여 학습을 한다. 그리고 무성음에서는 3프레임에 해당하는 FFT 진폭성분을 이동평균을 한다. 각각의 저역용 TDNN, 중역용 TDNN 및 고역용 TDNN으로부터의 출력을 합성하여 최종 FFT 진폭성분을 구한다. 그러나 위상성분은 진폭성분으로부터 직접 구한다. 마지막으로 역 고속 푸리에 변환(inverse fast Fourier transform : IFFT)을 사용하여 강조된 음성신호를 구한다.

III. 제안한 시간지연 신경회로망

본 논문에서 제안한 TDNN의 구조는 그림 2와 같이 나타내며 역전파알고리즘을 사용하여 학습한다. 입력 층의 가로축 방향은 n 프레임을 가진 시간을 나타내며, 여기에서 시간 축 방향의 최소 단위를 1프레임이라고 한다. 그림은 입력층의 4프레임의 유닛은 은닉층 제1층의 1프레임분의 유닛에 결합된다. 은닉층 제1층의 6프레임의 유닛은 은닉층 제2층의 1프레임의 유닛과 결합된다. 은닉층 제2층의 각 가로 1열의 유닛은 출력층의 각 유닛에 대응하여 접속되어 있다. 이러한 구조를 가짐으로써 신호의 시간변화 패턴을 반영할 수 있는 네트워크가 구축 가능하다. 따라서 본 논문에서는 21개의 FFT 진폭성분의 시간계열들은 n 프레임을 가진 입력층에 입력된다. 그 후에 입력층의 4 프레임은 첫 번째 중간층의 프레임에 연결된다. 30 유닛을 가진 첫 번째 중간층의 각 6프레임은 두 번째 중간층의 프레임에 연결된다. 그리고 21 유닛을 가진 두 번째 중간층의 각 프레임은 출력층에 연결된다.

본 실험에서는 FFT에 의해 구해진 진폭성분 중, 0~20샘플(0 kHz~1.2 kHz)은 저역(BPF1)부의 입력신호로, 21~41샘플(1.3 kHz~2.5 kHz)은 중역(BPF2)부의 입력신호로, 42~63샘플(2.6 kHz~3.9 kHz)은 고역(BPF3)부의 입력신호로 분할되어 입력되어 TDNN에 의하여 학습된다. TDNN의 입력신호에는 잡음이 중첩된 음성신호로부터 구해진 FFT 진폭성분이 부여되며 학습신호에는 잡음

을 부가하지 않은 음성신호로부터 구해진 FFT 성분을 부여하여 1프레임마다 학습을 한다. 본 실험에서는 제안한 TDNN들이 다음과 같은 4종류의 네트워크를 사용하여 학습되었다. (1) $SNR_{IN}(Input\,SNR) = 20\,dB$, (2) $SNR_{IN} = 15\,dB$, (3) $SNR_{IN} = 10\,dB$, (4) $SNR_{IN} = 5\,dB$. 학습의 실행에 필요한 각 TDNN의 여러 학습조건으로, 학습계수 α는 0.2, 가속도계수 β는 0.6, 초기하중은 -0.12 ~ 0.12의 난수, 입력의 실효값은 1.0으로 하여 TDNN을 학습시켰다. 본 실험에서는 최대 학습 횟수를 오차변화가 거의 없어지는 15,000회로 하였다.

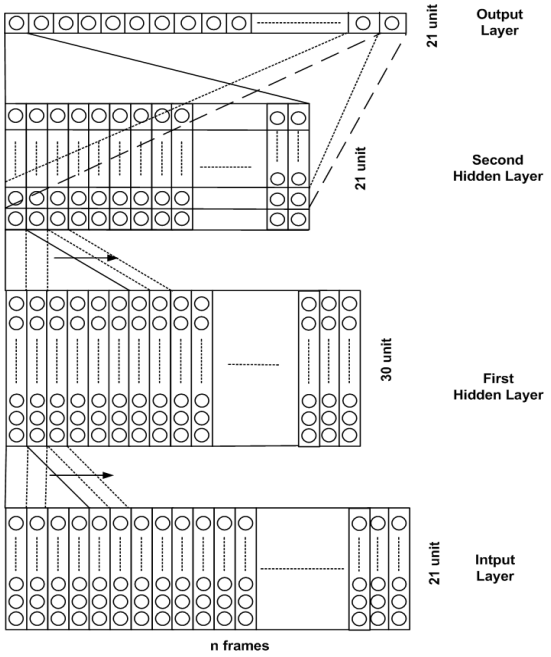

그림 2 제안한 저역용, 중역용 및 고역용 TDNN의 구조

V. 실험 결과

본 논문은 TDNN을 사용하여 음성신호를 강조하는 것을 목적으로 하여, 각 음성 데이터에 대한 음성강조 실험결과에 대해서 기술한다. 본 실험에서는 SNR을 20 dB ~ 5 dB의 환경 하에서 실시하여 본 방법의 유효성을 시간영역의 평가척도인 SNR_{out}(Output SNR)을 사용하여 본 방법의 유효성을 확인한다. 본 시스템의 성능평가를 위하여, Aurora2 데이터베이스의 테스트셋 A, B, C로부터 잡음이 중첩된 음성데이터들이 임의적으로 선택되었다. 제안한 시스템은 정상잡음인 백색잡음(white noise)에 대하여 TDNN에 의한 방법과 비교되었다. 그림 3은 백색잡음에 대하여 다양한 잡음레벨들($Input\ SNR = 20\,dB \sim 0\,dB$)을 사용하여, 20개의 문장에 대한 SNR_{out}의 평균값을 나타내었다. 그림 3의 백색잡음에 대하여, 잡음이 중첩된 음성신호(Original noisy speech)와 비교하였을 때, 각 대역별 TDNN을 사용하지 않은 경우(without TDNN)의 SNR_{out} 최대 개선값은 약 7dB, 본 방법은 약 8.5dB 개선되었다. 따라서 그림에 나타낸 것과 같이 제안한 시스템은 잡음레벨이 낮았을 때보다 잡음레벨이 높았을 때에 양호한 개선결과를 보였다.

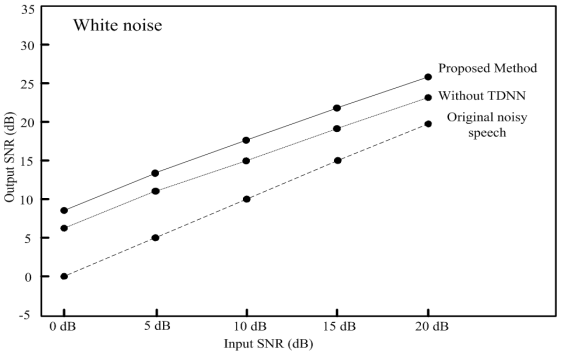

그림 3. 백색잡음 부가 시의 제안한 방식의 성능비교

VI. 결론

본 논문에서는 TDNN을 사용하여 잡음을 제거하는 시스템을 제안하여, 이것이 SNR에서 유효하다는 것을 백색잡음에 대해서 실험적으로 증명하였다. 따라서 제안한 시스템은 유성부 및 무성부에 대하여 각각 저역, 중역, 고역으로 분리된 시간지연신경회로망에 의하여 잡음이 제거됨을 확인할 수 있었다. 더욱이 각 대역별 TDNN을 사용하지 않은 경우와 비교하여도 본 방법이 유효하다는 것을 확인할 수 있었다. 결론적으로 본 연구에서는 입력 SNR이 5dB 정도의 조건에서도 충분히 잡음 제거 효과가 높다는 것을 확인하였다. 특히 저역부에 잡음이 집중한 유색잡음에 대해서도 저역부의 FFT 진폭성분을 복원하여 잡음을 제거할 수 있었다.

이상과 같이, 음성신호의 잡음제거를 위해서 TDNN에 의한 본 방식이 백색잡음에 대해서 효과적이라는 것을 실험적으로 확인하였지만, 향후의 연구과제로서는 다양한 유색잡음에 의해서 열화된 음성에 대해서도 더욱 강화하는 방법의 검토가 필요하다고 생각된다. 또한 신경회로망의 입력수가 많아짐에 따라 계산량이 증가하는 문제를 개선할 필요가 있으며, 입력샘플수를 증가시켰을 때에 학습능력을 향상시키기 위한 신경회로망의

학습조건을 변경시켜 학습시킬 필요가 있다고 본다.

이상으로, 본 논문에서 제안한 잡음에 강인한 잡음억제 시스템의 성과는 다양한 잡음 하에서의 잡음억제 및 음성강조에 도움이 될 것으로 생각된다.

참고문헌

[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] 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.

[3] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement", IEEE Trans. Speech and Audio Processing, Vol. 3, No. 6, pp. 433-438, 1995.

[4] 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.

[5] Widrow, B.; John, R.; Glover, J. R.; McCool, J. M.; Kaunitz, J.; Williams, C. S.; Hearn, R. H.; Zeidler, J. R.; Dong, E.; Goodlin, R. C.: Adaptive noise cancelling: Principles and applications. Proc. IEEE, 63, pp. 1692-1716, 1975.

[6] A. Ishida and H. Gobata, "Speech/Non-speech Discrimination under Real Life Environments". J. Acoust. Soc. Japan, Vol. 47, No. 12, pp. 911-917, 1991.

[7] Y. Wu and Y. Li, "Robust speech/non-speech detection in adverse conditions using the fuzzy polarity correlation method", IEEE International Conference on Systems, Man, and Cybernetics, Vol. 4, pp. 2935-2939, 2000.

[8] 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.