음성의 특징벡터를 사용한 정규화 인식수법

최재승*

*신라대학교 전자공학과

Normalized Recognition Method using Characteristic Vector of Speech Signal

Jae-Seung Choi*

*Department of Electronic Engineering, Silla University

E-mail : jschoi@silla.ac.kr

요 약

본 논문에서는 음성의 특징벡터를 추출하여 음성인식을 위한 인식 알고리즘을 제안한다. 본 논문에서 제안하는 방법은 사람의 음성을 정규화하여 시간지연신경회로망을 사용하여 음성인식을 하는 인식 알고리즘이다. 본 논문에서는 시간지연신경회로망을 이용하여 입력되는 음성정보를 일정시간 동안 학습시킨 후에 새로이 입력되는 정보를 인식하는 수법이다. 본 실험에서는 음성인식률에 의하여 본 알고리즘의 유효성을 확인한다.

키워드

Time Delay Neural network, Speech Recognition, FFT.

Ⅰ. 서 론

최근 디지털 장치 및 컴퓨터 등의 각종 정보기기의 보급에 의하여 디지털 컴퓨터의 응용 기술 및 디지털 신호처리 기술이 급격히 발전하게 되었다. 이에 따라 음성인식 및 화자인식 분야에서도 급속도로 컴퓨터의 응용기술을 이용한 연구가 다수 이루어지고 있다[1, 2].

특히 이러한 연구 중에서 신경회로망(Neural Network; NN)[2-5]의 오차역전파학습법(Back-propagation algorithm; BP)[5]을 응용하여, 이 신경회로망의 구조에 시간적 변화를 추가한 시간지연신경회로망(Time-Delay Neural Network; TDNN)[6]이 연구되고 있다. 따라서 본 논문에서는 이러한 TDNN을 이용하여 모음 및 자음에 대한 음성인식의 연구를 목적으로 하여 연구를 진행한다. 본 논문에서는 제안한 음성인식 알고리즘을 평가하기 위하여 음성인식률을 적용하여 화자인식이 효과적인 것을 실험으로 확인한다.

Ⅱ. 음성 특징 추출 방법 및 실험조건

본 실험에서는 샘플링 주파수 8 kHz의 이산시간신호를 128샘플(16 ms)의 프레임으로 분리하여 각 프레임의 샘플값을 해밍창를 통과시킨 후에 고속 푸리에 변환(Fast Fourier Transform, FFT)에 의해서 구해지는 FFT에 의한 FFT 켑스트럼 변환(FFT→log| |→IFFT)을 한다. 구해진 FFT 켑스트럼을 켑스트럼창에 통과시킴으로써 켑스트럼의 저역부의 12개의 켑스트럼 데이터를 구한다.

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어 숫자로 구성된 Aurora2 데이터베이스(Database, DB)[7] 및 일본 음성정보처리개발협회에서 배포한 연구용 연속음성 데이터베이스의 총 2종류를 사용하였다. Aurora2 DB의 모든 음성데이터는 ETSI (European Telecommunications Standards Institute)로부터 배포되었으며, 총 8440개의 숫자

로 구성된 테스트 셋 A, B, C의 음성데이터를 사용하였다. 일본 연구용 연속음성 데이터베이스는 총 26종류의 셋으로 구성되었으며 총 22,000 문장으로 구성되어 있다. 본 실험에서는 2가지 데이터베이스 중에서 임의적으로 30문장을 선택하였으며, 20문장은 신경회로망의 학습 데이터로 사용하며 나머지 데이터는 평가용으로 사용하였다.

음성신호 등을 입력으로 하여 신경회로망에 학습하는 경우, 신호의 시간변화가 중요한 요소가 되는 경우가 있다. 이러한 것을 고려한 신경회로망으로써 TDNN이 고안되었으며 본 논문에서는 TDNN을 사용한다.

Ⅲ. 제안한 화자인식 알고리즘

본 논문에서는 음성을 사용한 음성인식에 시간지연신경회로망을 사용한 음성인식의 식별방법을 제안한다. 본 논문에서 제안하는 화자인식 알고리즘의 처리 과정을 그림 1과 같이 나타낸다. 제안하는 음성인식 과정은 크게 음성신호의 전처리 과정, 유성음 및 무성음의 분리 과정, 유성음 및 무성음의 특징 추출 과정, 신경회로망에 의한 분류기, 음성인식 과정 등의 단계로 분류한다.

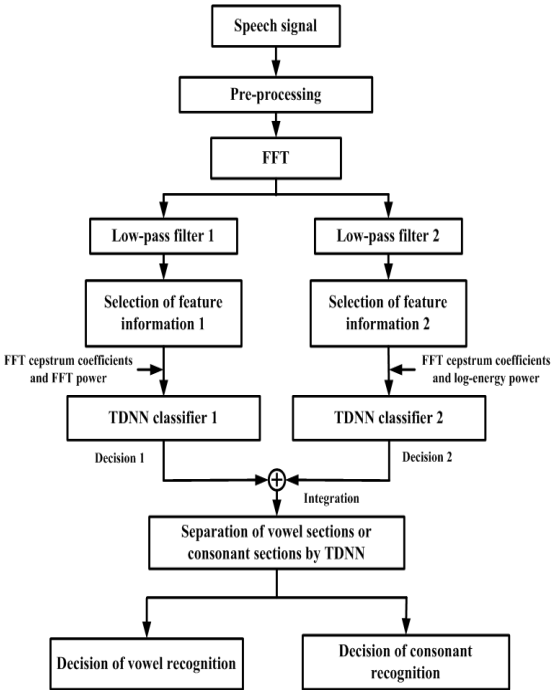

그림 1. 제안한 음성인식 알고리즘

그림 1의 제안한 음성인식 시스템에서의 시간지연신경회로망은 유성음 및 무성음 구간에 대하여 각각 입력층의 유닛수는 12개의 켑스트럼 및 1개의 전력 스펙트럼의 총 13개를 시간지연신경회로망에의 입력으로 한다. 중간층 유닛은 30개, 출력층은 2개의 유닛으로 구성된다. 따라서 학습을 통해 기억된 각 가중치들을 사용하여 새로운 화자의 음성이 입력될 경우 기억된 가중치들을 가지고 새로운 각 음성에 해당하는 음성에 대하여 유성음인가 아닌가를 인식하게 된다.

Ⅳ. 실험 결과

본 논문에서 제안한 시스템은 임의적으로 각 음성에 의한 특정 단어를 데이터베이스로부터 선택하여 음성인식 실험을 수행하여 음성인식률에 의하여 인식 성능을 평가한다. 본 실험에서는 시간지연신경회로망의 학습을 통해 기억된 각 가중치들은 13차의 새로운 음성에 대한 Cepstrum 값이 시간지연신경회로망의 입력으로 들어올 경우, 기억된 가중치들을 가지고 연산을 한 후에 각 음성에 대한 인식율을 구하게 된다.

본 논문에서의 음성 인식률은 발성음성의 전체 개수에 대하여 각 프레임에서 시간지연신경회로망의 출력값이 정확하게 검출되는 각 프레임 비율로 정의한다. 이 음성 인식률은 시간지연신경회로망의 학습이 완료된 후의 출력 가중치의 결합계수를 사용하여 각 프레임에서 인식율의 정확도를 측정하였다. 본 실험에서는 전체 프레임에 해당하는 문턱값을 90%로 설정하여 각 프레임에서 이 문턱값 이상이면 해당 음성으로 인식하게 되며 이는 우수한 인식 성능을 가지는 값을 실험적으로 선정한 것이다. 실험 수행 시 각각의 실험에 대하여 충분한 검증을 위하여 각각의 인식방법에 대해 총 10번 실험결과의 평균치를 사용하여 인식률을 산출하였다.

표 1은 FFT cepstrum 계수를 11차, 12차, 13차로 하였을 경우의 인식률을 나타낸다. 표 1에서 알 수 있듯이 FFT cepstrum 계수를 11차로 하였을 경우 89.6%, 12차로 하였을 경우 90.9%, 13차로 하였을 경우 90.1%의 인식률을 구하였다. 표 1의 실험 결과에서 알 수 있듯이 12차의 인식률이 가장 양호하였으며, 따라서 본 논문에서는 특징 파라미터로 12차의 FFT cepstrum 계수를 사용하였다.

표 1. FFT cepstrum 계수에 의한 인식률

FFT 켑스트럼의 계수	인식률[%]
11차	89.6%
12차	90.9%
13차	90.1%

V. 결론

본 논문에서는 기초적인 음성인식의 성능개선을 위하여 음성정보에 대한 변화에 따라 학습이 가능한 인간 뇌구조의 모델을 모의한 시간지연신경회로망을 사용하여 음성 인식률을 향상시키는 방법을 제안하였다. 제안한 시간지연신경회로망은 음성인식의 성능을 개선하기 위하여 오차 역전파 학습 알고리즘을 이용하여 네트워크를 학습시켰다. 제안한 인식 알고리즘은 발성음성의 유성음 구간 및 무성음 구간을 검출하고 검출된 음성구간에 대하여 FFT cepstrum 계수의 특징데이터를 추출한 후 이 특징데이터를 시간지연신경회로망에 적용시켜 음성을 인식하는 방법이다.
제안한 음성인식방법은 이러한 조건 하에서 학습을 한 후에 음성인식 실험을 통하여 각 음성의 인식 성능을 확인하였다.

참고문헌

[1] J. He, L. Liu, and G. Palm, “On the use of residual cepstrum in speech recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 5-8, 1996.

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

[3] 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.

[4] W. G. Knecht, M. E. Schenkel, G. S. Moschytz, “Neural network filters for speech enhancement”, Transactions on Speech and Audio Processing, Vol. 3, No. 6, pp. 433-438, 1995.

[5] A. V. Ooyen and B. Nienhuis, "Improving the convergence of the back-propagation algorithm", Neural Networks, vol. 5, no. 3, pp. 465-471, 1992.

[6] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition using time-delay neural networks," IEEE Trans. Acoust., Speech, Signal Processing, vol. 37, pp.328-339, mar. 1989.

[7] 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.