선형예측계수를 사용한 화자인식

최재승*, 정병구**
*신라대학교 전자공학과, **성화대학 항공전기전자과

Speaker Recognition using Linear Prediction Coefficient

Jae-Seung Choi*, Byeong-Goo Jeong**
*Department of Electronic Engineering, Silla University,
**Department of Aviation Electricity & Electronics, Sunghwa College

E-mail : *jschoi@silla.ac.kr, **jbg8917@hanmail.net

요 약

본 논문에서는 다층 퍼셉트론 신경회로망과 선형예측계수를 사용한 화자인식 알고리즘을 제안한다. 제안하는 화자인식 알고리즘은 입력받은 음성신호에 대해서 유성음 구간을 추출한다. 추출된 유성음 구간에 대하여 선형예측 분석에 의하여 화자의 특성을 가지고 있는 선형예측계수를 구한다. 구해진 선형예측계수를 분류하기 위하여 선형예측계수를 퍼셉트론 신경회로망의 입력으로 사용하여 네트워크의 학습을 수행한다. 본 실험에서는 선형예측계수와 신경회로망을 사용하여 본 화자인식 알고리즘이 유효하다는 것을 인식률을 통하여 확인한다.

키워드
Perceptron neural network, Linear prediction coefficient, Speaker recognition, Recognition rate.

Ⅰ. 서 론

근년 컴퓨터를 이용한 음성인식 및 화자인식은 신경회로망 연구의 발달로 활발히 연구가 진행되고 있다. 앞으로 음성은 인간이 컴퓨터에 입력하는 명령 수단으로서 인간이 가장 편리하게 이용하는 의사 전달 방식으로 될 것이다. 이와 같이 인간과 컴퓨터가 상호 대화가 가능하기 위해서는 컴퓨터가 인간이 말하는 음성을 인식해야 하며 이러한 것을 위한 연구가 신경회로망의 발달과 더불어 활발하게 이루어지고 있다. 이러한 신경회로망 모델은 Hopfield 모델, Kohenen 모델, 퍼셉트론, 다층 퍼셉트론 등의 모델이 있으며, 이들 중에서 오차역전파 학습 알고리즘(Back-Propagation Training Algorithm)[1, 2]에 의한 다층 퍼셉트론[3]을 사용하여 화자 인식이 연구되고 있다. 이 오차역전파 학습 알고리즘은 다층 퍼셉트론에 있어서 일반화된 학습방법으로서, 패턴 인식에 있어서 상당히 강력한 학습 알고리즘이라는 것이 다수의 연구에 의하여 증명되고 있다[1, 2]. 현재도 새로운 가능성을 제시하고 있는 신경회로망을 사용한 음성 인식의 연구가 공학 분야에서 계속 진행되고 있다[4, 5].

본 논문에서는 신경회로망을 이용하여 미래의 음성관련 컴퓨터가 특정 사람이 발성하는 음성을 인식하기 위한 기초연구를 수행하며, 여러 사람이 발성한 음성을 입력하여 각 개인이 가지고 있는 화자의 특징을 추출한 후에 이 특징 입력데이터를 신경회로망의 입력값으로 한다. 따라서 본 논문에서는 신경회로망을 오차가 거의 없어지는 일정 기간 동안 네트워크를 학습시킨 후에 신경회로망의 학습 데이터와는 다른 새로운 화자의 목소리를 신경회로망에 입력할 경우에 누가 발성한 음성인가를 판단하고 인식하는 화자인식 시스템을 제안한다.

II. 신경회로망의 구조 및 학습법

신경회로망은 인간의 두뇌를 모델링하기 위한 시도의 연구로부터 출발하였다. 1980년도부터 신경회로망이 가지고 있는 문제점을 극복할 수 있는 방법이 알려지기 시작하자 신경회로망을 이용하여 음성인식 등에 적용되기 시작하였다. 이 후 Rumelhart, McClelland, Hinton, Williams(1986) [1, 6] 등이 다층 신경회로망의 학습알고리즘인 오차 역전파 학습 알고리즘[1, 2]을 발표하기 시작하면서 신경회로망에 대한 관심이 고조되었다.

본 논문에서는 오차 역전파 학습 알고리즘을 사용한 신경회로망을 사용하였다. 오차 역전파 학습 알고리즘은 먼저 임의로 생성시킨 초기 가중치를 네트워크에 입력하여 생성된 출력값과 목표값과의 오차를 구한다. 이 후에 네트워크의 오차가 일정 범위 안에 들어오도록 감소시키면서 가중치를 최적값으로 수렴하도록 조절해 나간다. 따라서 이러한 오차 역전파 학습 알고리즘은 비선형적인 문제를 해결하는데 뛰어난 기능을 가지고 있다.

본 논문에서 사용한 신경회로망은 그림 1과 같은 입력층과 출력층 사이에 1개의 중간층을 가지는 Rumelhert[1]에 의해 제안된 2승오차 최소화의 학습 구조인 퍼셉트론(Perceptron)[3]형의 계층형 네트워크를 사용하며, 네트워크의 유닛 간은 입력층으로부터 출력층으로 향하는 순방향 결합을 가진다. 본 실험에서는 입력층이 12유닛, 중간층이 20유닛, 출력층이 4유닛을 가지는 구조를 가진다.

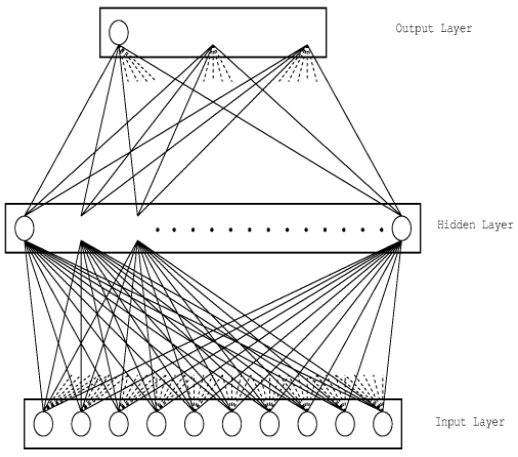

그림 1. 퍼셉트론 신경회로망

III. 화자인식 알고리즘

본 논문에서 제안하는 화자인식 알고리즘의 처리 과정을 그림 2와 같이 나타낸다. 본 논문에서 제안하는 화자인식 알고리즘은 입력으로 발성된 음성에 대하여 유성음 구간을 검출하는 구간분석 과정을 통해 발성음성 중에서 유성음 부분에 해당하는 구간을 검출한다. 검출된 유성음 부분에 대하여 특징추출을 위한 전처리 과정을 수행한 후에 선형예측(linear prediction) 분석을 통하여 신뢰성있는 특징 데이터를 추출한다. 본 논문에서는 학습과정이 간단하며 패턴분류 성능이 뛰어난 퍼셉트론형 신경회로망을 사용하여 추출된 특징 데이터들을 분류한다. 이러한 특성은 다양한 패턴을 가지고 있는 음성신호의 패턴분류와 인식에 적합하다. 분류된 결과값들은 후처리 과정에서 최종적인 매칭과정을 통하여 음성인식 성능을 평가한다.

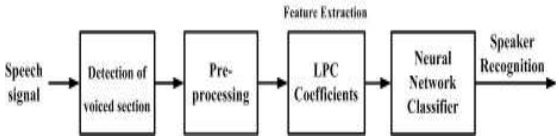

그림 2. 제안한 화자인식 알고리즘

음성 특징의 분류를 위한 신경회로망의 학습 데이터는 선형예측 분석을 통하여 추출된 12차의 선형예측 계수를 사용하였다. 각 프레임은 12차의 선형예측 계수로서 표현되며 각 프레임에서 추출된 계수값들은 학습을 위한 데이터로 사용한다.

본 실험은 Backpropagation 학습법을 이용하여 구현되었고, 신경회로망의 네트워크는 선형예측계수를 입력하기 위한 12개의 입력층 유닛, 20개의 중간층 유닛, 4개의 출력층 유닛으로 구성된 3층의 신경회로망으로 구성된다. 또한 신경회로망의 각 출력신호는 학습신호와 일치하도록 정확한 값을 취하도록 네트워크를 학습시킨다. 신경회로망의 학습 계수는 0.1, 가속도 계수는 0.6으로 하였으며, 최대 학습횟수는 10,000회로 하였다.

IV. 실험결과

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 영어숫자로 구성된 Aurora2 데이터베이스(Database, DB)[7]를 사용하였다. Aurora2 DB의 모든 음성데이터는 ETSI (European Telecommunications Standards Institute)로부터 배포되었으며, 테스트 셋 A, B, C의 음성데이터로 구성되어 있다[8]. Aurora2 데이터베이스는 남성화자 55명 및 여성화자 55명에 의해서 발성된 음성을 녹음한 총 8440개의 숫자로 구성된 테스트 셋 A, B, C의 음성데이터를 사용하였다. 본 실험에서 사용한 학습데이터는 4명의 화자가 발성한 4개의 문장에 대하여 각각 4번씩 발성한 전체 64개의 문장을 사용하였으며, 신경회로망의 학습용으로는 40개의 문장을, 테스트용으로는 24개의 문장을 사용하였다.

본 논문에서 제안한 시스템은 각 피실험자가 특

정의 동일한 단어를 발성하여 화자인식을 수행하여, 화자인식률에 의하여 인식 성능을 평가한다. 신경회로망의 출력 값은 목표 값과 비교하여 가중치를 수정하여 에러 값을 최소로 만든다. 목표 값은 화자가 4명이기 때문에, 화자 1인 경우에는 (1.0, -1.0, -1.0, -1.0)을, 화자 2인 경우에는 (-1.0, 1.0, -1.0, -1.0)을, 화자 3인 경우에는 (-1.0, -1.0, 1.0, -1.0)을, 화자 4인 경우에는 (-1.0, -1.0, -1.0, 1.0)을 각각 목표 값으로 설정하였다. 본 실험에서는 화자인식 전에 미리 기준패턴으로 화자에 해당하는 음성을 등록하고 발성하여 데이터베이스에 저장하는 형식으로 실험을 수행하였다. 따라서 이렇게 등록된 화자의 음성과 새로운 입력과의 화자매칭 과정을 통하여 최종적인 화자인식의 과정을 수행하게 된다. 실험 수행 시에 본 인식방법에 대해 총 10번 실험결과의 평균치를 사용하여 화자 인식률을 산출하였다. 따라서 화자 인식률은 발성음성의 전체 개수에 대하여 12차의 Cepstrum을 입력값으로 갖는 신경회로망의 출력값의 비율로 정의하며, 이 인식률의 비율이 90% 이상이면 해당 화자의 음성으로 인식하게 된다. 본 실험결과는 40개의 음성을 신경회로망에 의하여 학습을 완료한 후에, 총 24개의 목소리를 테스트하였는데 그 중에서 20개의 목소리를 정확히 인식하여 83.3%의 인식률을 보였다.

V. 결론

본 논문에서는 기초적인 화자종속 음성인식의 성능개선을 위하여 오차역전파알고리즘에 의한 신경회로망을 사용하여 화자 인식율을 향상시키는 방법을 제안하였다. 화자 인식을 위한 파라미터로서 선형예측계수를 사용하였으며, 이들 계수들을 신경회로망의 입력값으로 사용하였다. 제안하는 알고리즘은 발성음성의 음성구간을 검출하고 검출된 음성구간에 대하여 선형예측분석을 수행하여 선형예측 계수의 특징 데이터를 추출한 후 이 특징 데이터를 신경회로망에 적용시켜 화자를 인식하는 방법이다. 제안한 인식방법은 실험을 통하여 인식성능을 확인하였다. 향후 연구 과제로는 좀 더 많은 어휘의 인식이 가능한 화자독립 알고리즘을 연구할 예정이다.

참고문헌

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

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

[3] 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.

[4] S. Tamura, M. Nakamura, “Improvements to the noise reduction neural network”, 1990 International Conference on Acoustics, Speech, and Signal Processing, pp. 825-828, 1990.

[5] 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.

[6] L. Tan, P.C. Ching, L.W. Chan, “Recurrent neural networks for speech modeling and speech recognition”, International Conference on Acoustics, Speech, and Signal Processing, vol.5, pp. 3319 - 3322, 1995.

[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.

[8] R.G. Leonard, “A database for speaker independent digit recognition,” IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.328-331, Mar 1984.