LPC 켈스트럼 계수와 신경회로망을 사용한 화자인식

최재승*

Speaker Recognition using LPC cepstrum Coefficients and Neural Network

Jae-Seung Choi*

요 약

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

ABSTRACT

This paper proposes a speaker recognition algorithm using a perceptron neural network and LPC (Linear Predictive Coding) cepstrum coefficients. The proposed algorithm first detects the voiced sections at each frame. Then, the LPC cepstrum coefficients which have speaker characteristics are obtained by the linear predictive analysis for the detected voiced sections. To classify the obtained LPC cepstrum coefficients, a neural network is trained using the LPC cepstrum coefficients. In this experiment, the performance of the proposed algorithm was evaluated using the speech recognition rates based on the LPC cepstrum coefficients and the neural network.

키워드

퍼셉트론 신경회로망, 선형예측분석, 선형예측부호 켈스트럼 계수, 화자인식

Key word

Perceptron neural network, Linear predictive analysis, LPC cepstrum coefficient, Speaker recognition.

* 정 회 원 : 신라대학교 전자공학과(jschoi@silla.ac.kr)

접수일자 : 2011. 07. 14

심사완료일자 : 2011. 08. 26

I. 서 론

최근 음성신호를 이용하여 화자의 음성에 포함되어 있는 개인의 특징을 식별할 수 있는 정보를 바탕으로 하여 화자정보를 추출하는 화자인식 기술의 연구가 진행되고 있다[1, 2]. 앞으로 컴퓨터라는 의사전달 장치가 인간과 상호 대화가 가능하기 위해서는 컴퓨터라는 장치가 여러 사람들이 말하는 음성을 인식할 필요에 따라서, 이러한 것을 위한 연구가 신경회로망의 발달과 함께 현재 활발하게 연구가 진행되고 있다. 이러한 신경회로망 모델에는 대표적인 Hopfield 모델, 다층 퍼셉트론 등의 모델이 있으며[3], 본 논문에서는 특히 이들 중에서 오차 역전파 학습 알고리즘[4]에 의한 다층 퍼셉트론[5]을 사용하여 기초적인 화자 인식에 관한 연구를 진행한다. 이 오차역전파 학습 알고리즘은 다층 퍼셉트론에 있어서 보편화된 학습 알고리즘으로서, 음성에 관련한 패턴 인식 분야에 있어서 상당히 유용한 학습 알고리즘이라는 것이 여러 연구에 의하여 시험되어 증명되고 있다[3, 4, 5]. 현재 공학의 응용 분야에서 새로운 가능성을 제시하고 있는 신경회로망을 사용한 음성 인식의 연구가 진행되고 있다[6].

참고문헌[1]에서는 VQ (Vector Quantization) 및 HMM (Hidden Markov Model)에 기초하여 선형예측부호화 (Linear Predictive Coding, LPC) 분석에 의한 LPC 계수를 사용하여 연속음성 및 고립숫자의 화자인식을 수행하고 있다. 참고문헌[2]에서는 모음인식을 위하여 LPC 계수와 신경회로망을 사용하여 화자인식을 하고 있다. 참고문헌[1]과 [2]의 연구와는 다르게 본 논문에서는 각 프레임에서 음성구간을 효과적으로 검출하여 신경회로망에 정확한 특징정보를 입력하여 알고리즘을 설계하며, 이러한 기존 연구에 더하여 신경회로망의 구조 및 학습 방법 등을 개량하여 숫자음의 화자로 구성된 데이터베이스를 사용하여 화자인식을 수행한다.

본 논문에서는 신경회로망을 이용하여 미래의 음성 컴퓨터가 어떤 사람이 발성하는가를 인식하기 위한 기초연구를 수행하며, 여러 사람이 발성한 음성을 입력하여 각 개인이 가지고 있는 화자의 특징을 추출한 후에 이 특징 입력데이터를 신경회로망의 입력값으로 한다. 따라서 신경회로망의 학습 데이터와는 다른 새로운 화자의 목소리를 신경회로망에 입력할 경우에 어떤 사람이 발성한 목소리인가를 판단하고 인식하는 화자인식 시

스템을 제안한다.

II. 화자인식을 위한 신경회로망

2.1 일반적인 신경회로망

최근에 신경회로망이 음성인식 등에 있어서 상당히 효과적인 성능을 가지고 있다는 것과 화자인식 등에 있어서도 많은 성과 및 연구 방법이 알려지기 시작하자 신경회로망을 이용한 연구 등이 음성인식에 다수 적용되기 시작하였다. 또한 Rumelhart[4]에 의해서 제안된 2층 오차 최소화 학습 알고리즘을 다층 신경회로망의 학습에 확장시킨 오차 역전파 학습 알고리즘에 적용한 논문이 발표되면서 신경회로망에 대한 관심이 높아지기 시작하였다. 이 오차 역전파 학습 알고리즘은 먼저 랜덤하게 생성시킨 초기 가중치를 신경회로망의 네트워크에 입력하여 결과로 생성된 출력값과 희망하는 목표값과의 오차를 구한다. 이 후에 네트워크의 오차를 일정 범위 안에 들어오도록 감소시키면서 반복적인 학습을 통하여 가중치를 최적값으로 조절해 나간다.

2.2 제안한 신경회로망의 구성 및 학습 방법

본 논문에서 제안한 신경회로망은 그림 1과 같은 입력층과 출력층 사이에 1개의 중간층을 가지는 퍼셉트론[5]형의 네트워크를 사용하며, 네트워크의 유닛 간은 입력층으로부터 출력층으로 향하는 결합 구조를 가진다. 본 실험에서는 오차역전파 알고리즘 학습법을 이용하여 구현하였다. 화자인식기로 사용한 신경회로망의 구조는 14개의 입력층 유닛과 30개의 중간층 유닛수를 갖는 네트워크이다. 또한 출력층은 3명의 화자로 분류하기 위하여 3개로 구성하였다. 음성 특징의 분류를 위하여 14개의 LPC cepstrum 계수를 신경회로망에 입력함으로써 각 출력신호는 학습신호와 일치한 정확한 값을 취하도록 네트워크를 학습시킨다. 신경회로망의 출력값은 희망하는 목표값과 비교하여 가중치를 수정하여 에러 값을 최소로 만든다. 목표값은 화자가 3명이기 때문에, 화자 A가 선택될 경우에는 (+1.0, -1.0, -1.0)을, 화자 B가 선택될 경우에는 (-1.0, +1.0, -1.0)을, 화자 C가 선택될 경우에는 (-1.0, -1.0, +1.0)을 목표값으로 설정한다. 따라서 화자인식 알고리즘을 테스트할 경우에 학습을 통해 기억된 각 가중치들을 사용하여 새로운 화

자가 입력될 경우 학습에 의해 기억된 가중치들을 가지고 새로운 화자를 인식하게 된다. 신경회로망의 학습 계수는 $\alpha=0.2$, 가속도 계수는 $\beta=0.7$ 으로 하였으며, 최대 학습횟수는 12,000회로 하였다. 본 실험에서는 음성 데이터에 대하여 -1.0~+1.0의 사이에 해당하는 가중치로 정규화한다.

그림 1. 퍼셉트론 신경회로망
Fig. 1 Perceptron neural network.

III. 화자인식 알고리즘

본 논문에서 제안하는 화자인식 알고리즘의 처리 과정을 그림 2와 같이 나타낸다. 본 논문에서 제안하는 알고리즘은 입력된 음성에 대하여 전처리 과정을 수행한 후, 유성음 구간을 검출하는 구간분석 과정을 통해 음성 신호의 각 프레임에서 유성음 부분에 해당하는 구간을 검출한다. 검출된 유성음 부분에 대하여 신경회로망에 입력하기 위하여 선형예측 분석[7]을 통하여 특징을 추출한다. 본 논문에서는 패턴분류 성능이 우수한 퍼셉트론형 신경회로망을 사용하였다. 이러한 신경회로망은 여러 화자 특성 및 패턴을 가지고 있는 음성 신호의 패턴분류와 음성인식에 적합한 복잡한 패턴분류기로 적합하다고 본다. 신경회로망에 의해서 분류된 결과값들은 최종적인 화자인식의 결정과정을 통하여 인식성능을 비교함으로써 제안하는 알고리즘의 성능을 평가한다.

그림 3은 전처리 과정을 거친 음성신호의 세부적인 특징을 추출하기 위하여 음성에 대한 선형예측 분석 과정을 거친 후에 LPC cepstrum 계수를 구하는 LPC cepstrum 특징 추출의 블록 다이어그램을 나타낸다. 음성 특징의 분류를 위한 신경회로망의 학습 데이터는 14차의 LPC cepstrum 계수를 사용하였다. 각 프레임은 14차의 LPC cepstrum 계수로서 표현되며 각 프레임에서 추출된 계수값들은 학습을 위한 데이터로 사용한다.

그림 2. 제안한 화자인식 알고리즘
Fig. 2 Proposed speaker recognition algorithm.

그림 3. LPC cepstrum 추출의 블록다이어그램.
Fig. 3 Block diagram for LPC cepstrum extraction.

본 실험에서 사용한 음성신호는 8 kHz의 샘플링 주파수를 가진 환경에서 녹음된 연결된 영어숫자로 구성된 Aurora2 데이터베이스(Database, DB)[8]를 사용하였다. Aurora2 DB의 모든 음성데이터는 ETSI (European Telecommunications Standards Institute)로부터 배포되었으며, 테스트 셋 A, B, C의 음성데이터로 구성되어 있다. 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 데이터는 3명의 화자가 발성한 전체 64개의 문장을 Aurora2 데이터베이스 중에서 임의적으로 선택하였으며, 신경회로망의 학습용으로 40문장을, 테스트용으로 24문장을 사용하였다. 음성 데이터는 약 25프레임에서 50프레임 정도를 나누워지며, 1 프레임은 256 샘플로 구성된다.

3.1 화자인식을 위한 전처리 과정

전처리 과정은 음성 신호를 표본화하여 프레임처리 하는 과정, 표본화 처리된 입력 음성신호에서 유성음만을 추출하는 과정, 추출된 유성음 신호에서 선형예측 분석에 의한 선형예측부호화 cepstrum 계수를 추출하는 과정으로 구분된다.

1) 샘플링 분석 및 유성음 구간 검출

본 논문에서는 전체 음성신호에 대해서 구간별로 분석하기 위하여 식 (1)과 같은 해밍창을 사용하였다. 여기에서 N 은 분석구간 중의 표본수를 나타낸다.

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (1)$$

$$(n = 0, 1, \dots, N-1)$$

음성검출과정에서 보편적으로 사용되는 단시간 에너지 측정방법은 분석구간 단위로 수행하며, 해당 분석구간에서 에너지 측정값을 구하여 이 측정값 및 문턱값을 비교하여 해당 분석구간이 음성신호인지 아닌가를 판별한다. 따라서 본 논문에서 사용하는 에너지 측정방법은 전체 프레임의 평균 진폭값을 구하여 일정구간내의 진폭값이 에너지의 문턱값을 넘으면 음성구간으로 간주하는 방법을 적용한다. 본 논문에서는 해밍창을 통과한 음성신호는 각 프레임의 실효값 R_f 가 문턱값 $R_m/3$ 보다 큰 경우에는 유성음 구간으로 판별된다. 여기에서 R_f 는 각 프레임에서 구해진 실효값을 나타낸다. 본 실험에서는 처음의 약 5프레임에서 각 문장의 평균 실효값 R_m 을 실험적으로 구하였다.

2) 선형예측 분석

선형예측분석은 일반적으로 음성신호의 표본값 사이에는 상당한 상관관계가 있다는 가정을 이용하여 음성의 특징 추출을 위한 예측부호화가 실시되어진다[7]. 이러한 가정 하에서 선형예측오차의 2승 평균값을 최소로 하도록 저역에 해당하는 LPC cepstrum 계수를 결정한다. 또한 LPC cepstrum 계수는 동일한 음성이라 할지라도 화자의 개인의 감정 및 주변의 잡음환경에 따라 cepstrum의 값이 다양한 변화를 보여줄 수 있다. 따라서 본 논문에서는 이러한 cepstrum의 애매함을 개선하기 위하여 신경회로망의 이론을 사용하였다. 그러나 본 논문

에서는 개인의 감정상태 및 잡음환경 하에서의 연구는 향후 연구에서 구체적으로 다루고자 한다.

3.2 화자인식 알고리즘의 구현

본 논문에서 제안한 시스템은 임의적으로 각 화자에 의한 특정 단어를 선택하여 화자인식 실험을 수행하여, 화자인식률에 의하여 인식 성능을 평가한다. 본 논문에서는 신경회로망의 각 화자의 입력 특징 데이터들이 최종적인 화자인식의 결과로 판정되기 위하여, 화자인식 전에 화자의 음성을 미리 참조패턴으로 등록하여 데이터베이스에 저장하여 최종적인 화자인식의 결과를 판단하게 된다. 본 논문에서의 화자 인식률은 발성음성의 전체 개수에 대하여 신경회로망의 출력값이 정확하게 검출된 비율로 정의한다. 본 실험에서는 문턱값을 90%로 설정하여 이 문턱값 이상이면 해당 화자의 음성으로 인식하게 되며 이는 우수한 인식 성능을 가지는 값을 실험적으로 선정된 것이다. 실험 수행 시 각각의 실험에 대하여 충분한 검증을 위하여 각각의 인식방법에 대해 총 10번 실험결과와 평균치를 사용하여 인식률을 산출하였다.

IV. 실험결과 및 고찰

본 논문에서는 화자인식 알고리즘 구현을 목표로 하여, 제안한 알고리즘을 이용하여 각 화자를 인식하는 실험을 수행하였다. 실험환경은 Intel(R) Core(TM)2 Quad 2.4GHz CPU와 3.25GB RAM이 장착된 IBM 호환 컴퓨터에서 Windows XP OS를 사용하여 Microsoft VC++ 6.0으로 구현하였다. 음성은 III장에서 기술한 Aurora2 DB의 음성신호를 프로그램의 입력신호로 사용한다.

LPC cepstrum 계수를 신경회로망의 입력으로 사용하여 신경회로망의 학습을 완료한 후에 신경회로망에 의해 얻어진 인식결과를 표 1, 2, 3, 4에 나타내었다. 표 1은 신경회로망의 학습계수를 변경하여 실험을 하였으며, 표 2는 신경회로망의 학습 횟수를 변경하여 화자인식 실험을 한 결과를 나타내고 있다. 표 2에서 학습횟수가 12,000회이었을 때 인식률이 우수하였기 때문에 본 실험에서는 12,000회를 최적으로 판단하였다. 표 3은 LPC cepstrum 계수를 13차, 14차, 15차로 하였을 경우의 인식률을 나타낸다.

표 1. 학습계수 변경에 의한 인식률
Table. 1 Recognition rates by modification of training coefficients.

학습계수의 변경	$\alpha = 0.1,$ $\beta = 0.7$	$\alpha = 0.2,$ $\beta = 0.7$	$\alpha = 0.3,$ $\beta = 0.7$
인식률[%]	89.9%	91.7%	89.2%

표 2. 학습횟수에 따른 인식률
Table. 2 Recognition rates according to iteration numbers.

학습횟수	8,000	10,000	12,000
인식률[%]	87.8%	89.9%	91.7%

표 3. LPC cepstrum 계수에 의한 인식률
Table. 3 Recognition rates by LPC cepstrum coefficients.

LPC cepstrum 계수	13차	14차	15차
인식률[%]	88.3%	91.7%	89.8%

표 4. 제안한 방법과 기존방법의 성능비교
Table. 4 Comparisons of the proposed method and conventional method.

LPC cepstrum 계수	Proposed method	Conventional method
인식률[%]	91.7%	74.3%

표 3에서 알 수 있듯이 LPC cepstrum 계수를 13차로 하였을 경우 88.3%, 14차로 하였을 경우 91.7%, 15차로 하였을 경우 89.8%의 인식률을 구하였다. 표 3의 실험 결과에서 알 수 있듯이 14차의 인식률이 가장 양호하였으며, 따라서 본 논문에서는 특징 파라미터로 14차의 LPC cepstrum 계수를 사용하였다.

표 4는 본 논문에서 제안한 방법과 참고문헌[2]에서 제안한 기존방법과의 인식률의 비교를 나타내고 있다. 제안한 방법에서는 LPC cepstrum 계수를 14차로 하였고, 참고문헌[2]의 방법에서는 12차의 LPC cepstrum 계수를 사용하였기 때문에 동일한 조건에서 비교하기는 어렵다. 제안한 방법은 인식률이 91.7%인 반면에 참고문헌[2]의 방법[표 5 참조]에서는 3화자에 대한 평

균 인식률이 약 74.3%인 점을 고려한다면 본 논문에서 제안한 방법이 기존 방법보다 인식률이 향상되었음을 알 수 있다. 그러나 제안한 방법을 음성강조 등의 응용 시스템에 적용하기 위하여 화자인식률의 개선이 필요하다고 판단되며 이러한 점들을 향후의 연구과제로 한다. 인식률 실험과는 별도로 처리속도에 대한 성능평가를 실시하였는데, 14-30-3로 구성된 네트워크에 대하여 12,000회의 학습을 하였을 경우에 약 200초 정도의 시간이 소요되었다. 따라서 신경회로망의 학습 처리시간에 대해서는 충분히 성능개선이 되었다고 판단된다.

이상으로 신경회로망의 학습에 의한 인식률이 조금 낮게 나타난 것은 신경회로망의 학습 음성 데이터 및 화자수가 충분히 많지 않았기 때문이라고 생각된다. 그러나 91% 이상의 인식률을 구할 수 있었던 것은 LPC cepstrum 계수에 의한 각 화자에 대한 음성의 개인 특징 정보를 신경회로망에 의하여 어느 정도 분류할 수 있었기 때문이라고 판단된다. 따라서 이들의 인식률 및 처리속도에 대한 학습결과는 화자를 식별하는데 어느 정도 충분한 결과를 구했음을 나타낸다. 그러나 91.7%의 인식률과는 반대로 약 8.3% 정도의 오인식률이 발생하는 것은 화자인식에 필요한 입력데이터의 세분화, 중간층 유닛수의 증가, 학습계수 조정 등의 처리가 필요하다고 판단되므로 이러한 세부적인 내용들을 향후의 연구과제로 수행하여 보다 완성된 연구를 진행하고자 한다.

V. 결 론

본 논문에서는 기초적인 화자종속 음성인식의 성능개선을 위하여 오차역전파알고리즘에 의한 신경회로망을 사용하여 화자 인식률을 향상시키는 방법을 제안하였다. 제안한 알고리즘은 발생음성의 음성구간을 검출하고 검출된 음성구간에 대하여 선형예측분석을 수행하여 LPC cepstrum 계수의 특징데이터를 추출한 후 이 특징데이터를 신경회로망에 적용시켜 화자를 인식하는 방법이다.

본 논문에서는 신경회로망에 의해 학습 시 학습계수, 학습횟수, LPC cepstrum 계수를 변경하여 실험 하였으며, 그 결과는 입력층 유닛수 14개, 출력층 유닛수를 4개

로 하고, 중간층의 신경 유닛수가 30개일 때 가장 좋은 결과를 얻었으며, 학습계수는 0.2, 가속도 계수는 0.7, 최대 학습횟수는 12,000회가 가장 적절하였다. 제안한 화자 인식방법은 이러한 조건 하에서 학습을 한 후에 화자 인식 실험을 통하여 각 개인의 화자인식 성능을 확인하였다. 향후 연구 과제로는 좀 더 많은 음성 데이터 및 화자에 대하여 인식이 가능한 화자독립 알고리즘을 연구할 예정이다.

참고문헌

- [1] A. Revathi, Y. Venkataramani, "Speaker independent continuous speech and isolated digit recognition using VQ and HMM", 2011 International Conference on Communications and Signal Processing, pp. 198-202, 2011.
- [2] K. Kuah, M. Bodruzzaman, S. Zein-Sabatto, "A neural network-based text independent voice recognition system", Proceedings of the 1994 IEEE Southeastcon 'Creative Technology Transfer - A Global Affair', pp. 131-135, 1994.
- [3] B. Lu, J. J. Xu, "Research on Isolated Word Speech Recognition Based on Biomimetic Pattern Recognition", 2009 International Conference on Artificial Intelligence and Computational Intelligence, Vol. 2, pp. 436-439, 2009.
- [4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors", Nature, Vol. 323, pp. 533-536, 1986.
- [5] S. K. Pal, S. Mitra, "Multilayer perceptron, fuzzy sets, and classification", IEEE Transaction on Neural Networks, Vol. 3, No. 5, pp. 683-697, 1992.
- [6] 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.
- [7] P. B. Patil, "Multilayered network for LPC based speech recognition", IEEE Transactions on Consumer Electronics, Vol. 44, No. 2, pp. 435-438, 1998.

- [8] 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.

저자소개

최재승(Jae-Seung Choi)

1989년 조선대학교 전자공학과
공학사

1995년 일본 오사카시립대학
전자정보공학부 공학석사

1999년 일본 오사카시립대학 정보통신공학과
공학박사

2000년~2001년 일본 마쯔시다 전기산업주식회사
AVC사 전임연구원

2002년~2007 경북대 디지털기술연구소 책임연구원

2007년~현재 신라대학교 전자공학과 교수

※ 관심분야: 음성신호처리, 적응신호처리, 잡음제거,
신경회로망, 디지털통신, 멀티미디어 등