신경회로망을 이용한 화자종속 음성인식 성능에 관한 연구

A study of speaker dependent speech recognition using neural network

윤 지 원 , 이 종 수
연세대학교 대학원 기계공학과, 연세대학교 기계공학부

Ji Won Yoon, Jongsoo Lee
School of Mechanical Engineering, Yonsei University
E-mail : jwyoon@yonsei.ac.kr

ABSTRACT

본 연구는 화자종속 소어휘 음성인식의 성능을 개선하는 데 그 목적이 있다. 인식에 사용될 음성의 특징을 얻기 위해 Winer 필터와 LPC&Cepstrum을 이용하여 프레임 당 12차 패턴을 추출하였다. 추출된 특징패턴을 인식하는 인식부는 특히 소어휘 음성인식에 우수한 성능을 보이는 기존의 역전파 신경회로망(Backpropagation Neural Network)에 인식율 개선을 위하여 퍼지추론시스템을 결합한 형태로 구현되었다. 실험결과 신경망만을 사용한 경우에 비하여 인식율이 향상됨을 연구하였다.

Key words : 음성인식알고리즘, 신경회로망, 퍼지추론시스템

I. 서 론

현대사회는 정보화 사회이다. 수많은 정보의 수단 중에서도 음성은 가장 보편적이면서도 편리한 수단이다. 그러한 특성들로 인해 최근에는 음성신호를 인식할 수 있는 알고리즘과 시스템의 개발이 활발하게 진행중이다.[1]

본 논문은 화자종속 소어휘 음성인식 알고리즘의 성능을 개선하고 최적화하는데 그 목적이 있다. 인식에 사용될 음성의 특징으로 Winer 필터와 LPC&Cepsturm을 이용하여 프레임 당 12차 패턴을 추출하였으며, 추출된 Cepstrum 계수들은 퍼지추론시스템(FIS, Fuzzy Inference System)을 통하여 가공되고 역전파 신경회로망(BPN, Backpropagation Neural Network)의 입력층 노드에 입력된다.

본 논문의 구성은 다음과 같다. 먼저, 본론의 2.1에서 전반적인 음성인식과정과 프레임 당 12차 Cepstrum 계수를 추출하는 방법에 대해 알아보고, 2.2에서는 실질적인 인식부를 구성하는 퍼지추론시스템과 역전파 신경회로망에 대해서 살펴본 후 2.3에서 구현된 인식알고리즘을 이용한 실험을 통하여 그 실용성을 검증한다. 마지막 결론에서는 구현된 알고리즘의 평가 및 향후과제를 논하고 끝을 맺는다.

II. 본 론

2.1 음성인식의 개요

음성인식 과정은 크게 음성신호의 전처리 과정, 음성의 특성 추출 과정, 유사도 측정에 의한 패턴인식과정의 세 단계로 나눌 수 있다. 그림 1은 음성인식 과정의 개략도이며 기준 모델 집합을 구성하고 유사도 측정의 대상이 되는 음성의 특징을 추출하는 과정이 그림 2에 묘사되어 있다.

본 논문에서는 여러 가지 음성의 특징 중에서도 음성의 발성기관을 하나의 필터로 가정하고, 그 특징계수를 그 음성의 특징벡터로 사용

하는 LPC&Cepstrum을 사용하였으며 각기 다른 음성신호의 길이를 30프레임으로 정규화하여 사용하였다.[2][3][4]

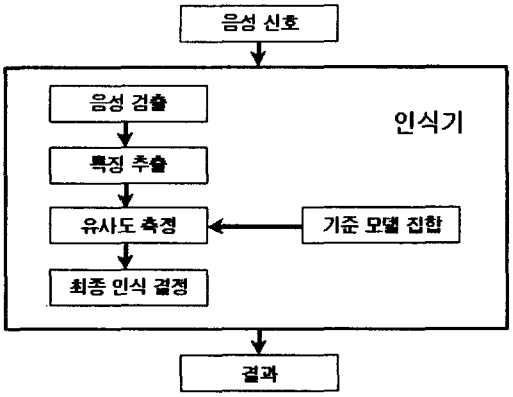

그림 1. 음성인식과정의 개략도

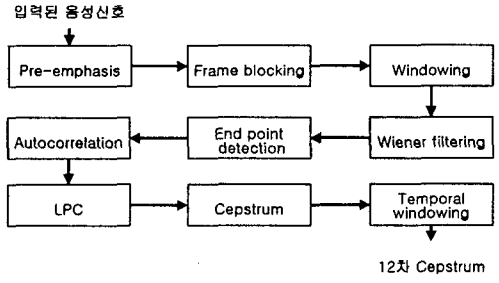

그림 2. 음성의 특징 추출과정

2.2 인식알고리즘의 구현

전술한 음성특징 추출방법을 통하여 추출된 12차 Cepstrum 계수의 예가 아래 그림 3에 나와있다. 본 연구에서는 최적화된 소어휘 음성인식 알고리즘 구현을 목표로 하므로 인식 대상어를 "앞으로", "뒤로", "서라", "왼쪽", "오른쪽"의 5개로 제한하였다.

그림 3. "앞으로"의 30프레임 정규화 12차 Cepstrum

그림 3에서 확인할 수 있듯이 동일한 음성이라 할지라도 화자의 상태나 기타 주변상황에 따라 Cepstrum의 값이 다양한 변화를 보여주고 있다. 이러한 다양성과 애매 모호함은 인식기의 인식능력을 저하시킨다. 본 연구에서는 이러한 Cepstrum의 애매 모호함을 처리하기 위하여 퍼지이론을 사용하였다.

2.2.1 퍼지추론시스템

퍼지 이론은 컴퓨터가 인공적 지능을 가지고 인간이 사용하는 수치는 물론 언어적으로 애매한 표현들을 처리할 수 있도록 한다. 즉, 기존의 논리체계는 0과 1의 개념이 확실한 반면 퍼지 이론은 어떤 집합의 소속정도를 0과 1사이의 값으로 표현함으로써 인간의 애매 모호한 상황도 표현할 수 있게 한 것이다.[5]

본 연구에서는 전술한 바와 같이 Cepstrum의 애매 모호한 성분을 제거하기 위하여 여러 가지 퍼지이론 중에서도 Mamdani가 제안한 추론방법인 직접법(min-max 추론법)을 사용하였다. Mamdani 직접법의 추론과정은 다음의 3단계로 나눌 수 있다.[5][6]

1단계 : 주어진 입력에 대해 각 규칙(Rule)의 전건부 멤버쉽 함수의 소속정도(적합도)를 구한다.
2단계 : 1단계에서 구한 소속정도를 기초로 각 규칙의 추론결과를 구한다.
3단계 : 각 규칙의 추론 결과에서 최종적인 추론 결과를 구한다.

이렇게 직접법을 통하여 퍼지함이 제거된 (Defuzzified) 12차 Cepstrum계수들은 역전파 신경망의 입력값으로 입력층 노드에 입력되어진다.

2.2.2 역전파신경회로망

신경회로망이란 생물학적인 뇌의 신경세포(Neuron)를 모델로 인공적으로 지능을 만드는 것이다. 즉, 인간의 뇌에 존재하는 신경세포와 이들의 연결 관계를 단순화시켜 수학적으로 모델링하여 두뇌가 갖는 지능적 기능을 인위적으로 구현한 것이다.[5][7]

본 논문에서는 여러 가지 신경망 중에서도 오류 역전파 규칙을 사용한 역전파신경회로망을 사용하였다. 오류 역전파 규칙(Error backpropagation)이란 먼저 초기의 연결 강도로 생성된 출력값과 목표값의 오차를 구한 후, 오차를 감소시키는 방향으로 연결 강도를 조절해 나가는 것을 말한다. 역전파 신경회로망은 특히 비선형 문제를 해결함에 있어 우수한 성능을 보인다.

전술한 방법에 의해 비퍼지화된 12차

Cepstrum을 입력값으로 갖는 신경망의 출력값은 아래의 식을 통해 판별되고 해당되는 음성으로 인식하게 된다.

$$(\frac{Y_k}{\sum_{k=1}^{K} Y_k}) \times 100 \geq x\% \qquad (1)$$

즉, 인식 가능한 음성의 수가 k 개일 때 특정 노드의 출력값이 전체 출력의 x % 이상이면 해당하는 음성으로 인식하게 된다. 본 연구에서는 x 의 값을 80으로 설정하였으며 이는 우수한 인식 성능을 가지는 값을 실험적으로 선정한 것이다.

2.2.3 음성인식을 위한 퍼지신경망

아래의 그림 4는 전술한 퍼지추론시스템과 역전파신경회로망등을 결합하여 최종적으로 구현된 인식알고리즘의 개략도를 보여준다.

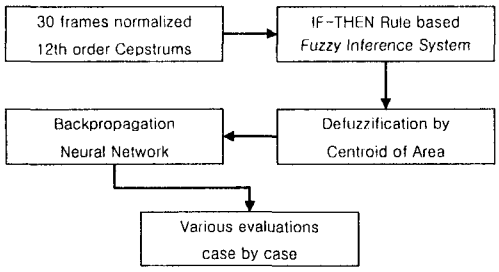

그림 4. 퍼지신경망을 응용한 음성인식알고리즘

2.3 실험 및 고찰

제안된 음성인식알고리즘을 이용하여 5개의 단어를 인식하는 실험을 수행하였다. 화자종속 음성인식성능의 개선이라는 목적을 고려하여 1사람의 화자가 발성한 목소리를 단어 당 30개씩 녹취하였으며, 녹취된 음성은 8kHz로 샘플링(sampling)하고 분해능(resolution) 16bit로 저장하였다. 그림 5부터 9까지는 저장된 음성신호로부터 추출된 언어별 30프레임 정규화 Cepstrum값들을 보여주고 있다.

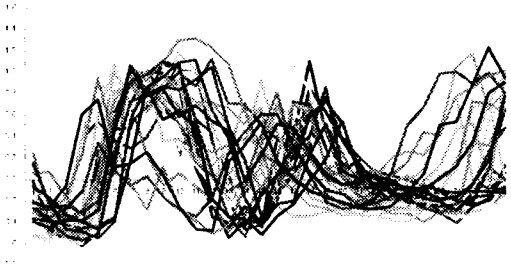

그림 5. "앞으로"의 30프레임 정규화 12차 Cepstrum

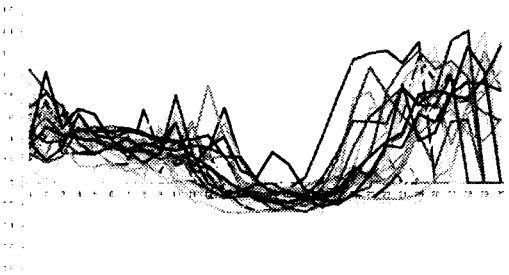

그림 6. "뒤로"의 30프레임 정규화 12차 Cepstrum

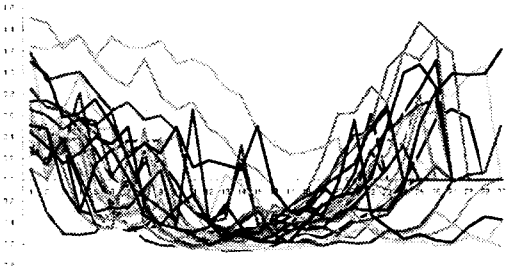

그림 7. "서라"의 30프레임 정규화 12차 Cepstrum

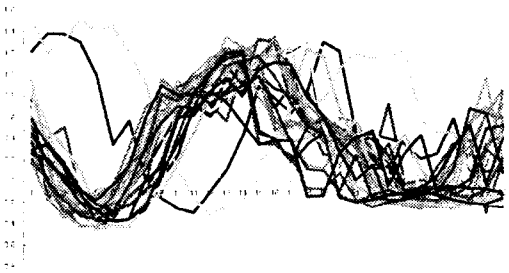

그림 8. "왼쪽"의 30프레임 정규화 12차 Cepstrum

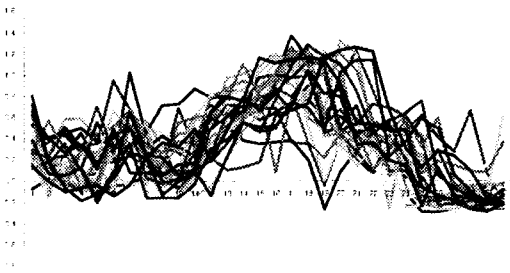

그림 9. "오른쪽"의 30프레임 정규화 12차 Cepstrum

퍼지추론시스템을 구축하는데 사용된 규칙(If-Then Rule)수는 146,688개 였으며, 비퍼지화 방법으로는 가장 유연한 결과를 보이는 식 (2)의 면적 도심법(Centroid of Area)를

사용하였다.

$$COA = \frac{\sum_{i=1}^{N} x_i \cdot \mu(x_i)}{\sum_{i=1}^{N} \mu(x_i)} \quad (2)$$

퍼지신경망의 성능을 비교평가하기 위하여 역전파신경망만을 사용한 음성인식의 실험조건 및 결과가 아래 표 1과 표 2에 나타나 있다. 역전파 신경망의 입력층 노드수는 정규화된 30프레임 켑스트럼을 입력하기 위하여 30개로 설정하였으며, 은닉층 노드수는 50개, 출력층 노드수는 5개로 설정하였다.

입력층 노드수	30
은닉층 노드수	50
출력층 노드수	5
Training pattern	150
Testing pattern	100
Iteration 수	50000
학 습 율	0.75

표 1. BPN의 구조 및 학습조건

단위 : %

앞으로	100 (20/20)
뒤로	95 (19/20)
서라	90 (18/20)
왼쪽	65 (13/20)
오른쪽	75 (15/20)
전체	**85 (85/100)**

표 2. BPN 사용 인식결과

다음의 표 3과 표 4는 본 연구에서 구현한 퍼지신경망을 사용한 음성인식의 실험조건 및 인식결과이다. 퍼지추론시스템을 통해 나온 COA가 5개의 노드로 구성된 입력층으로 입력되는 것을 제외하고는 역전파신경망만을 사용하였을 때와 동일한 조건에서 실험되었으며 결과적으로 BPN만을 사용하였을 때 보다 인식율이 높아졌음을 확인 할 수 있었다.

입력층 노드수	5
은닉층 노드수	50
출력층 노드수	5
Training pattern	150
Testing pattern	100
Iteration 수	50000
학 습 율	0.75

표 3. FNN의 구조 및 학습조건

단위 : %

앞으로	100 (20/20)
뒤로	95 (19/20)
서라	90 (18/20)
왼쪽	60 (12/20)
오른쪽	90 (18/20)
전체	**87 (87/100)**

표 4. FNN 사용 인식결과

실험결과 BPN과 FNN 모두 "왼쪽"의 인식율이 다른 단어들에 비하여 저조하다는 사실을 확인할 수 있었다. 이는 그림 8과 그림 5에서 확인할 수 있듯이 "왼쪽"의 Cepstrum 패턴과 "앞으로"의 패턴이 상당히 유사한 경향을 보이기 때문이라고 생각되어진다.

III. 결 론

본 연구에서는 소어휘 화자종속 음성인식의 성능개선을 위하여 역전파신경회로망의 애매모호한 입력값을 퍼지추론시스템으로 전처리하여 인식율을 향상시키는 방법을 연구하였다. 실험결과 처리되지 않은 입력값을 그대로 사용한 경우보다 인식율이 향상됨을 알 수 있었다.

향후 과제로는 좀 더 많은 어휘의 인식이 가능한 화자독립 알고리즘을 개발할 예정이다.

감사의 글 : 본 연구는 한국과학재단 지정 최적설계신기술연구센터의 연구비지원으로 수행되었습니다.

IV. 참고문헌

[1] 오영환, "음성언어정보처리", 홍릉과학출판사, 1998.
[2] 이행세, "음성인식기법", 청문각, 1999.
[3] Lawrence, Rabiner, Biing-Hwang Juang, "Fundamentals of speech recognition", Prentice Hall International Inc, 1993.
[4] 김정훈 외, "음성인식 기능을 탑재한 다기능 휠체어 시스템 설계 및 구현", 퍼지및지능시스템학회, 2002.
[5] 임영도 외, "퍼지•신경망•유전진화", 인솔미디어, 2002.
[6] 유동선 외, "기초퍼지이론", 교우사, 1998.
[7] 이현엽 외, "MATLAB을 이용한 퍼지-뉴로", 아진, 1999.