신경회로망을 이용한 연속음성중 키워드(Keyword)인식에 관한 연구

최 관 선, 한 민 홍
고려대학교 산업공학과

요 약 : 본 발표에서는 신경회로망을 이용하여 연속음성중에서 키워드를 인식하는 방법을 설명한다.

연속음성에서 파형소편(波形素片) 및 음절을 식별하는 휴리스틱 알고리즘을 개발하였고, 연속음성을 음절단위로 파형소편 스펙트럼분석(선형예측법)으로 특성치를 추출하였다.

음절의 특성치는 코호넨 신경회로망을 통하여 학습을 시켰으며, 연속음성중 키워드인식은 먼저 음절을 인식하여 단어를 찾고, 인식된 단어가 키워드와 일치하는가를 확인한다.

본 연구의 의의는 파형소편 및 음절식별 알고리즘을 통하여, 크기불변성(Scaling invariance), 시간불변성(Time warping 및 Time-shift invariance), 중복성제거의 문제점을 해결하였고, 신경회로망의 학습을 통하여 화자독립적인 연속음성인식시스템 구축의 기반을 확립한데 있다.

본 음성인식모델은 학교구내 전화번호 안내시스템으로 활용단계에 있으며 전화번호뿐만아니라 주소안내시스템으로도 활용될 예정이다. 또한 자동차 운전보조시스템 및 주행안내시스템의 음성명령에 응용될 수 있는데, 예로 음성명령은 "핸들 좌로 20도", "시청까지 주행", "시청 지도안내"등이 될 수 있다. 현재 자동차 운전보조시스템은 컴퓨터 화면상 모의동작시스템으로 운영되고 있다.

본 음성인식모델은 화자종속시 90%이상, 화자독립시 70%의 인식결과를 보였다.

I. 서론

음성인식을 구현하기 위해서는 음성신호로부터 특징되는 패턴을 안정적으로 추출할 수 있는 능력이 있어야 한다. 우선 음성신호의 시간축 왜곡현상(Time Warping)을 처리해야 하므로 TimeWarping Invariant해야 한다. 다음에 단어의 시작과 끝을 미리 알 수가 없으므로 인식단위구간중 어느 부분에 단어가 위치하더라도 찾아낼 수 있어야 하므로 Time Invariant해야 한다. 아울러 특징추출의 능력이 음성신호의 절대적 크기에 관계해야 하므로 Scale Invariant해야 한다. 아울러 특징추출 능력이 화자에 따라 변해서는 안되므로 Speaker Invariant해야 한다. 이런 Invariance 특성과 패턴인식 능력을 이용할 수 있다면 화자독립 연속음성인식을 실현할 수 있을 것이다[3].

기존 연구되어 온 음성인식알고리즘을 살펴보면 크게 템플릿기반(Template-Based Approach)으로서 DTW(Dynamic Time Warping), HMM(Hidden Markov Model), 지식기반시스템(Knowledge-Based Approach), 신경망(Connectionist Approach)으로 구분할 수 있다. DTW는

Dynamic Programming을 HMM은 확률추정(Stochastic Estimation)을 지식기반시스템은 인공지능을 이용한 추론(Inference)을 신경망은 패턴분류(Pattern Classifition)의 기능을 이용해 동일한 문제을 각기 다른 방법으로 풀고 있다고 볼 수 있다. 현재 가장 앞선 시스템이 HMM이라 할 수 있고 신경망 시스템은 아직 연구단계에 있다[3].

그동안 연구되어온 음성인식연구에서는 음성신호에서 특성치(parameter)를 추출할 경우 일정한 길이의 프레임단위(frame unit)로 나누어 특성치를 추출하였다. 이방법으로는 어느 시점에서 프레임을 끊느냐에 따라 그 프레임의 특성치가 달라지는 시간축상의 왜곡현상이 발생한다.
또한 단어간의 경계를 명확하게 인식할 수 없다.

본 연구에서는 음성신호의 시간축상 왜곡현상 및 동일파형의 중복성을 해결할 수 있는 새로운 특성치 추출방법으로 파형소편 추출알고리즘을 개발하고 음성인식의 화자독립을 위한 신경망의 학습방법에 대해서 연구하였다. 또한 연속음성중에서 음절을 추출하고 인식함으로써 키워드(Key Word)을 식별하는 알고리즘에 대해서 연구하였다.

II. 연속음성중 키워드(Key Word) 인식

2.1 음성특성치(parameter) 추출방법

그동안 연구되어온 음성인식연구에서는 음성신호에서 특성치(parameter)를 추출할 경우 일정한 길이의 프레임단위(frame unit)로 나누어 특성치를 추출하였다. 이방법으로는 어느 시점에서 프레임을 끊느냐에 따라 그 프레임의 특성치가 달라지는 시간축상의 왜곡현상이 발생한다.

본 연구에서는 이러한 시간축상 왜곡현상을 해결하기 위해 음성신호에서 파형소편을 추출를 하고 파형소편을 비교하는 방법을 개발하였다. 파형소편이란 파형부호화 용어로 음성신호는 몇가지 단위파형이 반복되어 구성되는데 이 단위파형을 말한다. 파형소편은 음성신호파형의 기본 구성요소로, 시작점과 끝점이 일관적이므로 시간축상 왜곡현상(Time Wraping)이 없어 진다. 음성신호 특성치를 추출하는데 반복되는 파형소편을 제거함으로써 인식시간을 단축하고 인식의 정확도를 향상시킬 수 있다.

음성신호는 PCM(Pulse Code Modulation)방식으로 샘플링 주파수를 10 KHZ, 양자화 정밀도(resolution)를 8 bit로 녹음한 후, 녹음된 음성신호를 다시 3개의 샘플데이타의 평균(Smoothing Method)한 값을 갖고 파형소편(波形素片)을 추출하였다.

다음은 파형소편 추출절차(procedure)를 설명한다. 추출절차에 사용된 용어정의로 block은 음파중 반복되는 단위구간, 파형소편을 칭한다. block size는 파형소편의 크기를 칭한다. block_high[i]는 i 파형소편의 시작점 진폭(amplitude)이다. block_inv[i]는 i 파형소편의 간격(interval)이다.

단계 1 : 유성음에 해당되는 파형소편 추출

1) 음성신호중 peak, valley detect

2) 인접신호들의 높이 비교
다음 block detect 조건에 만족한 시점의 신호를 block으로 한다.
전 peak(valley) 점과 현 valley(peak) 점으로 이루어지는 선분에서
- 전 peak(valley) 높이가 상역치(high threshold) 이상 인가?
- 현 valley(peak) 높이가 하역치(low threshold) 이하 인가?
- 경사도가 경사도역치(incline threshold) 이하인가?
- 크기가 크기역치(size threshold) 이상인가?
- 전 block과의 간격이 간격역치(interval threshold) 이상인가?

단계 2 : 음성시작점 추출
단계 1에서 추출된 첫번째 파형소편(block)부터 역방향으로 일정구간(크기 5) 음성 진폭의 분산을 계산한다. 분산이 분산역치(variance threshold) 이하로 일정 간격 이상 계속되면 최종 분산역치 이상인 시점을 음성시작점으로 한다.

본 연구에서 block detect를 위해 사용한 역치(threshold)값은 다음과 같다.
- 상역치(high threshold) : 130
- 하역치(low threshold) : 117
- 경사도역치(incline threshold) : 1
- 크기역치(size threshold) : 17
- 간격역치(interval threshold) : 20
- 분산역치(variance threshold) : 2.0

이 역치값들은 샘플링 레이트(sample rate) 및 양자화 정밀도(resolution)따라 적당히 조절되야 한다.

다음은 음성신호중에서 음절을 추출해 내는 방법을 설명한다. 음절추출은 파형소편(block)의 높이와 간격(interval)으로 추출한다.

음절추출 절차는 다음과 같다.

단계 1 : i 번째 block의 높이, block_high[i]와 간격, block_inv[i]을 구한다.
(i : block 번호)

단계 2 : 현 block과 이전 연속 N개의 block high의 차이, dif[j]와 누적차이, sum를 계산한다.
dif[j] = 1 if block_high[i-j] - block_high[i] > 2
dif[j] = -1 if block_high[i-j] - block_high[i] < -2
(i : 현 block 번호, j = 1 ~ N : 이전 block 번호)

$$sum = \sum_{j=1}^{N} dif[j]$$

단계 3-1 : 음절끝점 검사

a) 단음절인 경우

마지막 block이 음절끝점이 된다.

b) 음절이 2 이상인 경우

check_sign[i] = check_sign[i-1]+1 if sum > 5

check_sign1[i] = check_sign1[i-1]+1 if check_sign > 1

연속 block에서 음절끝점 if sum < -8 and check_sign1 > 5 and block size > 30

마지막 음절에서는 마지막 block이 음절끝점이 된다.

단계 3-2 : 단계 3-1에서 추출하지 못한 음절끝점 검사

value = block_high[i] x block_inv[i+1]

value가 local maximum 값을 갖는 block이 음절끝점에 해당한다.

음성신호의 특성치는 음절단위로 추출된다. 음성신호에서 추출된 음절은 음절의 시작부분(초성), 중간부분(중성), 끝부분(종성)의 각각 2개의 파형소편에 대해서 선형예측법(LPC : Linear Predictive Coding)으로 LPC계수를 산정하였다. 한음절은 120차수(order)의 벡터(vector)로 이루어진다. 각 파형소편은 해밍창(Hamming window)에 의해 창화(windowing)되었다. 해밍창(hamming window)는 아래식과 같다[2].

$$w(n) = 0.54 - 0.46 \cos\frac{2\pi n}{N}, \quad 0 \le n \le N-1$$

$$w(n) = 0 \qquad \text{otherwise}$$

여기서 N은 window 길이이다. 기존연구에서는 음성신호의 window 길이는 일반적으로 20-40ms 범위로 30ms를 주로 사용하였다. 이와같이 window 길이, N을 상수화 한다면 시간축상 이동(Time shift)으로 일관성있는 프레임(frame)의 특성치을 추출할 수 없다. 그러나 본 연구에서는 window 길이 N을 파형소편의 길이로 사용하였다. 동일파형소편들의 약간의 길이차이는 실지 추출된 특성치에는 거의 영향이 없어 무시할 수 있었다.

2.2 신경망을 이용한 음절특성치의 학습(Learning)

1980년대에 접어들면서 사람의 정보처리 능력이 음성, 화상, 제어분야에서 컴퓨터보다 탁월하다는 사실에 근거하여 연구되기 시작한 신경망은 사람의 정보처리 과정을 모델링하여, 간단하고 많은 처리요소들을 병렬로 상호 연결하여 학습을 통해 입력패턴에 내재하는 정보를 스스로 찾아내어 처리할 수 있다. 사람의 자연스러운 음성을 인식하기 위해서는 무엇보다도 매우 높은 계산속도가 요구되는데 신경망은 간단하고 많은 처리요소들을 병렬로 연결하여 높은 계산속도를 제공할 수 있다. 또한 신경회로망의 연결강도의 일부분이 여러가지의 에러에 의해서 훼손되어도 신경회로망은 특별한 오동작을 하지않으며, 이는 역으로 잡음이나 생각치 않은 요인에 의해 입력패턴이 변형될 경우에도 정보를 처리할 수 있다[3].

음성인식용 신경망은 동작특성에따라 정적(Static)인 신경망과 동적(Dynamic)인 신경망으로 구분할 수 있다. 또 한편 음성인식용 신경망을 학습(Learning)능력에 따라 지도학습(Supervised Learning)과 자율학습(Unsupervised Learning)신경망으로 구분하기도 한다. 정적인 신경망의 대표적인 것은 MLP(Multilayer Perceptron), SOFM(Self Organization Freature Map), ART가 있다. 동적신경망의 대표적인 것은 TDNN(Time Delay Neural Net)과 Recurrent Neural Net이 될 것이다. 아울러 지도학습의 대표적인 예는 MLP, Hopfield Net, Hamming Net이고 자율학습의 대표적인 예는 SOFM, ART, Darwin II 이다[3].

음성인식에서 화자독립문제는 신경망의 학습능력을 통한 일반화(Generalization) 기능으로 해결할 수 있다. 본 연구에서는 음성인식의 화자독립을 위하여 음절특성치를 SOFM 신경망인 코호넨 신경망(kohonen Neural Network)으로 학습시켰다. 코호넨 신경망 구조는 그림 1과 같다. 신경망은 2개의 층으로 이루어져 첫번째 층은 입력층(input layer)이고 두번째 층은 경쟁층인데 2차원의 격자(grid)로 되어있다. 모든 연결들은 첫번째 층에서 두번째 층의 방향으로 되어 있으며 두번째 층은 완전 연결(fully connected)되어 있다. 코호넨의 학습은 각 뉴런에서 연결강도 벡터와 입력벡터가 얼마나 가까운가를 계산한다. 그리고 각 뉴런들은 학습할 수 있는 특권을 받으려고 서로 경쟁하는데 거리가 가장 가까운 뉴런이 승리하게 된다. 이 승자 뉴런이 출력신호를 보낼 수 있는 유일한 뉴런이다. 또한 이 뉴런과 이와 인접한 이웃 뉴런들만이 제시된 입력벡터에 대하여 학습할 수 있다. 코호넨의 학습규칙은 다음 식으로 표현된다.

$$W_{new} = W_{old} + \alpha(X - W_{old})$$

여기에서 W_{old}는 조정되기 이전의 연결강도 벡터이며, W_{new}는 조정후의 새로운 연결강도 벡터이다. X는 입력벡터이며, α는 학습상수이다.

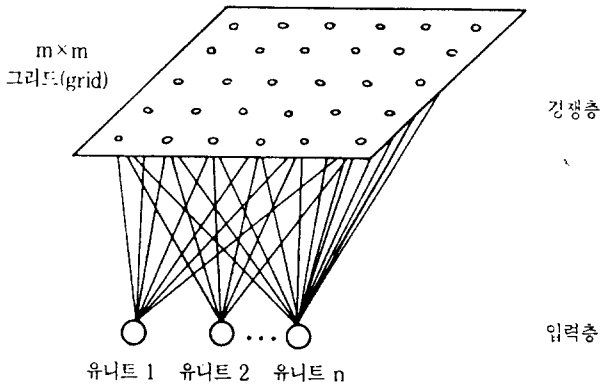

그림 1. 코호넨 네트워크

2.3 연속음성중 키워드(Key Word)인식

마이크를 통해 들어온 연속음성신호는 음절단위로 구분되어 특성치를 추출한다. 음절단위 특성치(입력벡터)는 코호넨 신경망을 통해 학습된 참조(reference) 음절(연결강도 벡터)들과 비교되고, 참조음절중 경쟁에서 승리한 음절(승리뉴런)이 해당음절로 인식하게 된다. 인식된 음절은 순차적으로 미리 정해놓은 단어집(Word File)에서 단어들의 음절과 비교하여 단어를 찾게 된다. 인식된 음절들의 순차적 결합이 미리 정해논 단어집(Word File)에 있는 단어라면 이를 키워드로(Key Word)로 인식하게 된다. 연속음성신호중에서 키워드가 아닌 음절은 인식시간과 조음현상에의한 인식정확도에는 영향을 주지만 실질적 키워드 인식에는 영향을 주지 않는다. 또한 단어의 순서도 키워드인식에는 영향을 주지 않는다.

III. 음성인식 사례연구

개발된 음성인식시스템은 자동 전화번호 및 주소안내시스템, 운전보조시스템, 주행안내시스템으로 응용되었다.

자동 전화번호 및 주소안내시스템은 사용자가 "홍길동 전화번호는 무엇", "전화번호 홍길동", "홍길동 전화번호","홍길동 주소는 무엇" 등으로 음성명령을 하였을 경우 "홍길동씨의 전화번호는 000-0000입니다", "홍길동씨의 주소는 안암동입니다"로 전화번호 및 주소안내를 해준다.

운전보조시스템은 음성을 이용한 운전보조시스템으로 "핸들 좌로 30도", "속도 60", "속도 후진 10", "정지", "경적"등 운전에 필요한 몇가지 명령어에 대해서 응답할 수 있다. 이 시스템은 현재 컴퓨터 화면상으로 모의되고 있다.

주행자동안내시스템은 GPS(Global Positioning System)장비를 이용한 주행중 운전자에게 도로안내를 해주는 시스템이다. 운전자가 "시청 주행", "시청 지도 안내", "주유소 위치"등 운전중 알고싶은 사항을 음성으로 알려주면 주행자동안내시스템은 적절한 정보를 제공해 준다.

IV. 결론

본 연구에서는 연속음성중에서 키워드(Key Word)을 인식하는 방법에 대해서 연구하였다. 음성인식시 파형소편 및 음절추출 알고리즘을 개발하여, 시간축상의 왜곡문제를 해결하였고, 음절특성치의 신경망학습을 통하여 화자독립방안을 제시하였다. 연속음성신호에서 음절단위로 음절을 인식하고 미리 정해진 단어집에서 단어를 찾아냄으로써 음성단어순서에 관계없이 키워드를 인식할 수 있었다.

본 음성인식시스템의 사례연구결과 화자종속인 경우 90%이상 인식결과를 보였고, 화자독립인경우 70%정도의 인식률을 보였다. 완전한 화자독립을 위해서는 추가연구가 필요하다.

참고문헌

[1] 동역메카트로닉스 연구소, "음성합성과 음성인식 시스템", 영진출판사, 1990

[2] Panos E. Papamichalis," Practical Approaches To Speech Coding", 1987

[3] 정 홍, "신경망을 이용한 음성인식", 전기공학회 논문지, 10권 2호, pp 49~59, 1992년 4월

[4] Lawrence R. Rabiner," A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, vol. 77, NO. 2, pp. 257-286, Feb. 1989

[5] Jay G. Wilpon, Lawrence R. Rabiner, Chin-Hui Lee, E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, no. 11, pp. 1870-1878, Nov. 1990