韓國 OA學會 論文誌
第5卷 第2號, 2000. 6.

2000-5-2-3-2

선형 변환망을 이용한 화자적응 음성인식

이 기 희*

Speaker Adaptation Using Linear Transformation Network in Speech Recognition

Kee Hee Lee*

요 약

본 논문에서는 불특정 화자의 음성에 대해서도 신뢰성 있는 인식이 이루어지도록 하는 음성인식 시스템을 구현하기 위한 화자적응 음성인식 기법을 제안한다. 제안한 화자적응 기법에 의한 음성인식 시스템은 표준화자의 음성특징을 1차 선형 변환망에 의해 새로운 화자의 음성특징에 선형적으로 적응하여 인식하며, 그 구성은 다층퍼셉트론을 퍼지 벡터양자화기로 사용하는 반연속 HMM을 기반으로 한다. 구현한 인식시스템은 그 성능을 확인하기 위해 고립단어 인식실험을 수행하였다. 그 결과, 화자적응 인식인 경우가 화자적응 수행하지 않은 시스템에 비해 인식률이 개선됨을 보였다.

Abstract

This paper describes an speaker-adaptive speech recognition system which make a reliable recognition of speech signal for new speakers. In the proposed method, an speech spectrum of new speaker is adapted to the reference speech spectrum by using parameters of a 1st linear transformation network at the front of phoneme classification neural network. And the recognition system is based on semicontinuous HMM(hidden markov model) which use the multilayer perceptron as a fuzzy vector quantizer. The experiments on the isolated word recognition are performed to show the recognition rate of the recognition system. In the case of speaker adaptation recognition, the recognition rate show significant improvement for the unadapted recognition system.

* 동서울대학 사무자동화과 조교수
** 본 논문은 동서울대학 산업기술 연구소의 지원으로 연구되었음.

Ⅰ. 서론

인간은 다양한 매체를 통하여 서로 정보를 교환하고 있다. 이와 같은 정보교환은 보다 편리하고 신속하게 하기 위하여 다양한 매체의 통합인 멀티미디어 환경 하에서 인간과 기계와의 인터페이스 기술이 중요하게 되고 있다. 인간이 사용하는 음성은 인간의 대화 수단 중에서 가장 자연스러운 의사 소통의 도구이다. 이 음성을 활용하여 컴퓨터를 중심으로 하는 기계와의 정보 교환 기술은 매우 중요한 것이다. 인간의 말은 성대의 울림에 의한 공기의 진동이 성도를 통해 입 밖으로 나옴으로써 생성된다. 이러한 말은 어떤 정보를 가지고 있으며, 말은 정보의 변별적 특성에 의해 표현된다. 이 변별적 특성은 입과 혀의 모양 및 위치에 따라 성도의 모양이 달라짐으로써 성대의 울림에 의해 발생한 공기 진동이 특정주파수에서 공진을 일으키거나 또는 공기의 진동을 막고 틔움으로써 발성한다. 이러한 인간의 말을 기계가 이해하도록 하는 음성인식 기술은 정보통신의 발전과 함께 그 필요성이 증대되고, 맨 머신 인터페이스 기술들 중에서도 핵심이 되는 기술이다. 음성은 인간이 비록 유사한 단어 음성을 발성하더라도 발성자의 심리적, 신체적 상태에 따라 많은 차이가 있다. 또한 주변환경에 따라 음성신호는 많은 변이와 변동을 지니고 있다. 이들을 극복할 수 있는 맨 머신 인터페이스를 위한 음성인식기를 개발하기 위하여 많은 과학자들이 꾸준히 연구를 수행해 오고 있다. 아직까지는 인간의 음성을 기계가 완벽하게 인식하는 시스템이 나오지 못하고 있다. 그러나 제한된 조건하에서는 인간의 음성을 이해하는 기술은 개발되어 상품화가 이루어지고 있다. 이 음성인식은 특정 화자의 음성에 대해서는 만족할 성능을 보이고 있지만 불특정 화자의 음성의 인식에는 성능이 떨어지고 있다. 따라서 음성인식기가 불특정 화자에 대해서도 신뢰성 있는 인식이 이루어지도록 하는 음성인식 시스템을 구현하기 위하여 화자적응화 기술에 대한 연구가 필요하다.

음성인식시스템에서 화자적응 기술은 인식기의 종류에 따라 여러 가지가 있으며, 대부분 HMM 인식시스템[1,2]에 기초를 두고 있다. HMM에 기초한 화자적응 기술은 이산 HMM인 경우에는 코드북 적응과 HMM 파라미터 적응[3]을 사용하여 원래의 HMM을 새로운 화자에 적응시키고, 연속 HMM인 경우에는 평균벡터 적응과 공분산행렬 적응기법[4] 등의 방법이 사용된다. 그리고 스펙트럼 사상을 통해 특징 벡터를 변환하는 방법[5], 신경망의 비선형 매핑에 기초한 LVQ를 이용한 방법[6] 등 다양한 적응 방법이 연구되고 있다. 이와 같이 음성인식 시스템의 화자적응기술에는 여러 가지 방법이 있다. 그러나 인식시스템의 표준패턴의 파라미터들에 새로운 화자들의 파라미터을 정규화시켜 적응인식을 수행한다는 기본적인 개념은 모두 동일하다.

본 논문에서는 화자적응 음성인식 시스템을 구현하여 인식성능을 향상시키는 인식기법을 제안한다. 인식시스템은 표준화자의 음성특징을 1차 선형 변환망에 의해 새로운 화자의 음성특징에 선형적으로 적응하여 음성을 인식하며, 그 구성은 전단에 음소분류신경망을 가지는 반연속 HMM을 기반으로 한다. 구현된 인식시스템의 성능은 단어음 음성데이터를 이용하여 실험을 수행하여 평가하고, 그 효율성을 보인다.

Ⅱ. HMM과 MLP를 이용한 음성인식 방법

1. HMM에 기반한 음성인식

HMM을 이용한 음성인식은 직관적인 알고리즘인 DTW[7]와는 달리 확률이론에 근거하고, 융통성이 많아 고립단어 인식에서부터 연속음성 인식에까지 광범위하게 사용되고 있다. 인식방법은 음성이 마코프 프로세스로 모델링될 수 있다는 가정 하에 훈련과 인식과정으로 구분되어 수행된다. 학습과정에서는 학습데이터의 관측심볼열 $O=(o_1, o_2, \cdots, o_T)$ (T는 관측심볼열의 길이)로부터 조건부 확률 λ가 최적화 기준에 맞도록 HMM λ의 파라미터를 추정하여 인식하고자 하는 단어의 모델을 결정한다. 인식과정에서는 각 HHM에 대한 음성입력 패턴의 조건부 확률 $P(O|\lambda)$를 구한 후, 조건부 확률이 최

대가 되는 HMM에 해당하는 단어를 인식된 음성으로 결정한다. HMM을 이용한 음성인식은 화자독립, 연속음성 인식에 많은 장점을 가지고 있으며, 계산량도 DTW에 비해 작다.

HMM은 상태천이확률과 출력확률로 표시되는 이중의 프로세서이다. 상태천이는 마코프 프로세서로 출력확률은 세 가지로 표시된다. 첫째, 벡터양자화를 통해 코드북의 코드워드로 나타낼 수 있다. 모든 가능한 음향적인 특성을 이 코드북에 이산확률밀도 함수로 나타낼 수 있다. 둘째, 연속확률밀도 함수로 표현할 수 있다. 이는 해당 단위에 종속적으로 해당 음성을 가지고 여러 개의 음성특징 벡터에 대한 평균과 분산을 구하기 때문이다. 셋째, 이것은 이산확률밀도와 연속확률밀도 함수를 혼합한 것이다. 확률분포가 이산적인 이산 HMM에는 음성특징벡터에 해당하는 관측심볼이 벡터양자화를 통해 가장 근접한 코드워드로 대표된다. 따라서 계산정도가 비교적 작고 간단하지만 벡터양자화에 의한 정보손실이 불가피하여 양자화 오류가 존재한다. 이산 HMM에서 발생되는 양자화 오류를 극복하기 위해 연속 HMM이 제안되었다. 이 모델은 추정할 파라미터가 너무 많기 때문에 추정을 위한 큰 데이터베이스화가 필요하고 많은 계산량을 요구하게 된다. 그러나 이산 HMM이 벡터양자화로 인한 왜곡을 수반하는 것에 반하여 음성의 특징벡터를 직접 사용하여 모델링하므로 인식성능이 이산 HMM에 비해 우수하다. 또한 연속 HMM은 초기 값에 민감한 경향이 단점으로 지적되고 있기 때문에 이산 HMM과 연속 HMM의 장점만을 선별해 만든 준연속 HMM가 있다. 이 모델은 벡터양자화시에 각 코드북에 존재할 확률 값으로 연속 HMM을 근사화한 방법이며, 많은 음성인식기에 이용되고 있다.

2. 다층신경망을 이용한 음성인식

신경망이란 상호 연결된 많은 수의 인공 뉴런들을 이용하여 생물학적인 시스템의 계산 능력을 모방하는 소프트웨어나 하드웨어로 구현된 계산 모델을 말한다. 인간의 뇌는 뉴런이라는 신경의 기본단위로 구성되어 있으며, 이들간의 상호 밀도 있는 연결 형태에 따라 지식을 암호화하거나 해독하게 된다. 신경회로망에서는 생물학적인 뉴런기능을 단순화시킨 인공 뉴런을 사용한다. 그리고 가중치를 갖는 연결선을 통해 상호 연결시켜 인간의 인지작용이나 학습과정을 수행하게 된다.

음성인식의 문제에 있어서 신경망은 학습에 의해 그 패턴에 대한 지식을 습득할 수 있으며, 표준 패턴의 음성특징과 비슷한 음성특징을 갖는 패턴들을 인식해 낼 수 있는 장점을 갖고 있다. 그리고 이들은 병렬처리를 수행함으로써 계산속도를 높일 수 있으며, 신경망에 내재된 정보가 신경망내의 모든 계산단위에 전파되기 때문에 구조내의 잡음이나 결함에 대해 신경회로망이 크게 영향을 받지 않으므로 신뢰도와 내고장성을 갖는다. 그리고 적응적 학습방법에 따라 연결가중치를 고정하지 않고 실시간에 적용되게 함으로써 인식률을 향상시킬 수 있다. 신경망을 학습하는 방법으로는 크게 자율학습과 교사학습으로 나눌 수 있다. 자율학습이란 입력패턴만 주어질 뿐, 출력패턴은 주어지지 않은 학습방법으로 코드북 작성 등에 사용되는 LVQ[6]등이 있다. 한편 교사학습은 입력과 출력패턴이 미리 주어지는 학습방법이다. 이 학습법은 대부분의 신경회로망에서 사용되며, 주로 음소인식이나 음소분류를 위한 경우에 사용되는 것으로 다층퍼셉트론이 있다. 이 다층퍼셉트론은 음성인식의 전처리 과정에서 음성에 존재하는 잡음이나 불필요한 특징의 제거, 고역주파수 강조 등의 기능을 위해 사용된다.

본 논문에서 구현한 기본 음성인식 시스템은 HMM의 전단에 다층퍼셉트론 이용하여 음소분류인식을 수행하도록 하고 있다. 이처럼 다층퍼셉트론을 이용한 음성인식 시스템에서는 다층퍼셉트론을 벡터양자화기 또는 확률추정기로서 사용될 수 있다. 벡터양자화기로 사용하는 방법에서는 프레임 단위의 음성신호를 다층퍼셉트론의 입력층에 넣고 출력층에서는 하나의 심볼을 얻는다. 각 프레임별로 구한 심볼열로부터 이산 HMM이 구성된다. 따라서 인식시스템은 간단한 반면에 훈련 데이타의 수가 많아야 하고 양자화 오류로 인해 인식률이 저하되는 단점을 가지고 있다. 반면, 다층퍼셉트론을 확률추정기로 사용하는 방법은 다층퍼셉트론의 출력 값을 관측확률의 추정 값으로 보고, 이로부터 HMM을 구성한다. 이 방법은 다층퍼셉트론의 학습이 어렵고, HMM의 모든 상태가 하나의 음소로 제한되는 단점이 있다. 이러한 단점을 감소시키는 방법으로 입력벡터에 대해 각 코드워드에 정합되는 정도를 나타내는 퍼지 스코어를 발생시키는 퍼지 벡터양자화기[8]가 있다. 본 논문에서도 다층퍼셉트론을 퍼지벡터 양자화기로 이용한다. 본 논문에서 구현한 인식시스템에서 음소분류신경망으로 사용한 다층퍼셉트론 학습은 오차역전파 알고리듬으로 이루어진다. 이때 식 1과 같이 입력패턴 x_i에 대해 i 층과 그 다음 j 층 사이의 가중치 w_{ji}를

선형 결합하여 얻어진 값에 식 2의 시그모이드 비선형 함수를 활성화 함수로 취하여 최종적인 출력값 o_j를 얻는다. 이러한 처리는 각각의 인공 뉴런에 대해 수행한다.

$$net_j = \sum_i w_{ji} o_i \quad (1)$$

$$o_j = \frac{1}{1+e^{-(net_j+\theta_j)}} \quad (2)$$

Ⅲ. 화자적응 인식 시스템

화자적응 인식을 위한 인식시스템의 구성은 다음의 그림 1과 같다. 인식시스템은 전처리과정을 통해 얻어진 음성신호를 LPC 분석으로 추출된 음성특징벡터를 입력받아 전단에 1차 선형 변환망을 가지는 음소분류신경망을 거쳐 HMM에 의해 화자적응 음성인식을 수행하게 된다. 이때 1차 선형 변환망은 새로운 화자의 음성 스펙트럼이 표준화자의 음성 스펙트럼에 적응시키는 역할을 수행하며, 음소분류신경망은 유사 음소간의 변별력을 높일 수 있도록 유사 음소그룹별로 나누고, 각 음소그룹별로 신경망이 구성되어 있다.

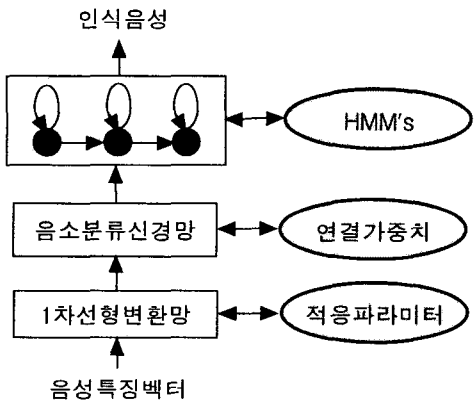

그림 1. 화자적응 인식시스템 블록도
Fig 1. Block diagram of speaker-adaptive recognition system.

1. 선형변환망과 음소분류신경망

화자적응 인식시스템에 사용한 음소분류신경망은 학습시간을 결정하는 학습횟수를 줄이고, 비슷한 특성을 갖는 음성과의 분별력을 갖게 하여 인식률을 높일 수 있도록 단어음에 포함되어 있는 음소들을 특성이 유사한 음소들로 묶어 그룹화하여 그룹별로 신경망을 학습하도록 하였다. 각 신경망은 그 그룹에 속하는 음소 데이터만을 대상으로 학습하는 방법을 사용하였다.

우리말의 음소는 조음적인 측면에서 공명음, 폐쇄음 또는 파열음, 마찰음, 그리고 파찰음의 네 가지로 분류하고 있다. 따라서 음소그룹신경망을 위한 각 음소의 그룹화는 인식 단어음에 포함된 음소들 중에서 우리말 음소 특징에 따라 모음에 속하는 음소 ㅏ, ㅓ, ㅗ, ㅜ, ㅚ, ㅔ, ㄹ* 의 그룹 G1, 비음에 속하는 음소 ㄴ, ㅁ, ㅇ, ㄴ*, ㅁ* 의 그룹 G2, 무성음에 속하는 음소 집합을 음소 발성 특징에 따라 구분하여 ㄱ, ㄷ, ㅂ, ㅌ, ㅍ, ㄱ* 의 음소 그룹 G3, 마찰음과 파찰음에 속하는 ㅅ, ㅈ, ㅊ, ㅎ 의 음소 그룹 G4로 구분하여 모두 4개의 음소그룹으로 분류하였다. 여기서 음소 ㄹ은 음절의 받침이 될 때 혀의 끝이 치경에 닿아 혀의 양변으로 공기의 흐름이 나누어지면서 발성이 된다. 이때는 모음과 비슷한 특성을 가지므로 모음 음소 그룹으로 분류하였다. 그리고 단모음에 속하는 음소 ㅐ 와 ㅔ 는 다차원 거리상의 척도가 매우 근접하여 음성 특성이 매우 유사하기 때문에 인식시에 ㅔ 와 동일한 모음으로 취급하여 인식하도록 하였다. 그리고 *로 표시된 음소는 음성의 종성부분에서 해당하는 음소를 나타낸다.

인식하려는 음성은 매 프로임마다 특징 파라미터를 추출하여 그 프래임이 어떤 음소그룹에 속하는 특징을 지닌 것인지를 먼저 각 그룹분류신경망을 사용하여 분류하게 된다. 따라서 음소분류신경망은 유사 음소별로 그룹화하여 학습하는 4개의 신경망 그룹과 이들 그룹을 분류하는 1개의 신경망 G5로 이루어진 모두 5개의 그룹 신경망으로 그림 2와 같이 구성하였다.

각각의 그룹신경망은 하나의 은닉층을 가지고 있는 다층퍼셉트론으로 구성되며, 오차 역전파 알고리듬에 의해 학습이 이루어지도록 하였다. 학습과정에서 수렴속도를 빠르게 하기 위하여 학습 가중치 조정시에 바로 이전의 가중치에 대한 일부를 현재의 가중치에 포함시키도록 식 3과 같이 모멘텀항 $\alpha()$를 사용하였다.

$$w_{ji}(t+1) = w_{ji}(t) + \eta \delta_j o_j + \alpha(w_{ji}(t) - w_{ji}(t-1))$$

(3)여기서 δ_j는 은닉층 노드 j 에서의 오차항이고, w_{ji}는 입력층 노드 j 와 은닉층의 노드 i 사이에 연결된 연결

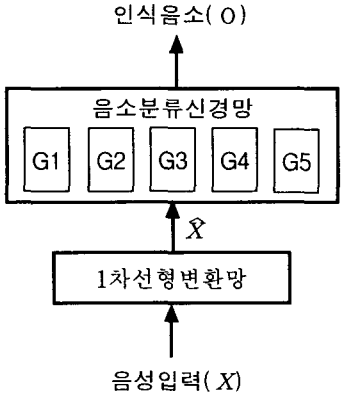

그림 2. 변환망을 사용한 음소분류신경망
Fig 2. Phoneme classification neural network using the transformation network.

가중치를 나타낸다. 그리고 화자적응 인식시에 변환망은 입력음성 X에 대해 음소분류신경망의 출력 값이 HMM의 관측확률을 증가시킬 수 있도록 식 4에 의해 변환하여 적응 인식하게 한다.

$$\hat{X} = AX^T + b \tag{4}$$

식에서 A와 b 는 적응 파라미터로서 음소분류신경망을 통해 오차를 역전파하여 식 5와 식 6에 의해 조정되어 진다.

$$a_{ij} = a_{ij} + \eta\delta_j w_{ji} c_j \tag{5}$$

$$b_i = b_i + \eta\delta_j w_{ji} \tag{6}$$

여기서 c_j는 음성 특징벡터 X를 이루는 LPC 켑스트럼 계수를 나타내며, 본 논문에서는 $1 \le j \le 16$이다.

2. 다층퍼셉트론을 이용한 HMM 인식

음소분류신경망을 이용한 화자적응 음성인식기의 HMM은 단순 좌우구조의 모델이다. 따라서 HMM 파라미터의 집합은 $\lambda = (\Pi, A, B)$로 표현되며, 이의 마코프 체인은 초기상태 확률벡터 $\Pi = (1, 0, 0, \cdots, 0)$와 상태천이 행렬 $A = [a_{ij}]\ (1 \le i, j \le N)$로 표현된다. 여기서 a_{ij}는 현재의 상태 i 에서 다음의 상태 j 로 천이할 확률로서 $\sum_j a_{ij} = 1 (1 \le i \le N)$의 조건을 만족한다. 또, 관측확률 행렬 $B = [b_j(O_t)]\ (1 \le j \le N,\ 1 \le t \le T)$로 나타내며, N은 상태 수이고 T는 입력음성 프레임의 길이이다. 음성인식 시스템에서 음소분류신경망은 시간 t에서 음성프레임의 특징벡터열 $X = (X_1, X_2, \cdots, X_T)$를 입력으로 하여 음소단위의 분류인식을 하도록 학습된 다층퍼셉트론 신경망이다. 이때 다층퍼셉트론은 입력 음성벡터열 $X = (X_1, X_2, \cdots, X_T)$를 각 코드워드에 정합되는 정도를 나타내는 퍼지 스코어를 발생시키는 퍼지 관측열 $O = (O_1, O_2, \cdots, O_T)$로변환하며, $O_t = (o_{t1}, o_{t2}, \cdots, o_{tM})$이다. 이 다층퍼셉트론의 출력 값을 관측확률의 추정 값으로 보고, 이로부터 HMM을 구성한다.

다층퍼셉트론을 확률추정기로 사용하는 음성인식 방법에서는 입력음성 벡터 X에 대해 음소 클래스 m에 대응하는 다층퍼셉트론의 출력노드의 값은 주어진 입력음성 X에 대한 사후확률 $P(m|X)$의 추정 값이 된다. 이와 같이 다층퍼셉트론을 확률추정기로 HMM과 결합하면 음소클래스 m 에서 입력음성 벡터 X의 확률 $p(X|m)$ 즉, 유사성은 식 7로 구할 수 있다.

$$p(X|m) = \frac{P(m|X)}{P(m)} p(X) \tag{7}$$

여기서 $P(m|X)$는 MLP의 출력값으로 X의 사후확률이고, $P(m)$는 음소 m의 확률로서 m의 상대 빈도가 되며, $p(X)$는 X의 확률이다. 이와 같이 MLP를 이용하여 입력 음성의 유사성을 구하면, 이를 기존의 HMM과 결합시킬 수 있다. 본 논문에서도 MLP를 FVQ로서 사용하였다. 따라서 다층퍼셉트론의 출력노드는 각 음소 m에 대응되며, 입력 X_t에 대한 각 MLP의 m번째 출력노드의 값 o_{tm}은 음소 m의 사후확률 $P(m|X_t)$의 추정 값이 된다. MLP를 FVQ로 이용한 HMM은 HMM λ의 각 상태 j에서 O_t의 관측 확률밀도 함수 $b_j(O_t)$는 다음 식 8로 표현할 수 있다.

$$\begin{aligned} b_j(O_t) &= P(O_t|q_t = s_j, \lambda) \\ &= \sum_{m=1}^{M} c_{jm} o_{tm}, \quad 1 \le j \le N \end{aligned} \tag{8}$$

여기서 q_t는 상태 레이블이고, M은 MLP의 출력 노드 수이다. 그리고 o_{tm}은 정규화한 출력값으로 그 합이 1이 되도록 제한한다. 식 5는 기존의 HMM에서 관측확

률 밀도함수 $b_j(\cdot)$를 다층퍼셉트론의 출력값 o_{tm}에 가중치 c_{jm}를 곱하여 구한다. 여기서 c_{jm}는 상태 j 에서 $m(1\le m\le M)$번째 가지의 가중치를 식 9으로 나타내며, 식 10을 만족해야 한다.

$$\sum_{m=1}^{M} c_{jm} = 1, \quad 1\le j\le N \tag{9}$$

$$c_{jm}\ge 0, \quad 1\le j\le N,\ 1\le m\le M \tag{10}$$

c_{jm}의 재추정식은

$$\hat{c}_{jm} = \frac{\sum_{t=1}^{T} P(O|q_t = s_j, h_t = m|\lambda)/P(O|\lambda)}{\sum_{t=1}^{T} P(O|q_t = s_j|\lambda)/P(O|\lambda)} \tag{11}$$

이다. 여기서 λ는 HMM이고, $\hat{c}_{jm}$는 재 추정된 값이며, q_t와 h_t는 시간 t 에서 O_t가 발생된 상태 및 다층신경망의 출력노드를 각각 나타낸다. 이와 같이 입력벡터열 X를 퍼지 관측열 O로 변환하고 상태 j의 새로운 관측확률 $b_j(O_t)$를 구하면, HMM의 파라미터 행렬 A와 B는 Baum-Welch 재추정식[9]을 이용하여 구할 수 있다. 이러한 구조의 인식시스템은 신경망의 변별력을 HMM에 부여하여 인식률을 개선한다. 그리고 하나의 다층퍼셉트론을 모든 HMM이 공유하여 사용할 수 있어 시스템이 간단해진다.

본 논문의 화자적응 인식은 학습과정과 인식과정으로 이루어지며, 학습과정은 HMM의 기준 모델을 훈련시키는 단계이다. 오차 역전파 알고리듬에 의해 학습되어진 다층퍼셉트론의 연결가중치를 이용하여 계산된 출력노드 값을 이용하여 HMM 재추정식에 의해 훈련과정을 통하여 HMM 파라미터를 추정하였다. 훈련과정을 통해 모든 기준 HMM을 만들어 두고, 인식과정에서는 학습된 각 단어 HMM λ_m에 음성 신호를 전처리하여 추출한 음성의 특징벡터를 입력벡터로 하여 비터비 스코어에 의해 관측확률 $P(O|\lambda)$를 구한다. 이 관측확률 중에서 가장 높은 확률에 해당하는 HMM의 단어모델을 인식된 단어로 판정한다.

Ⅳ. 실험결과 및 검토

화자적응 인식 시스템의 성능평가를 위한 인식실험은 음소분류신경망의 음소인식을 수행하여 성능을 확인하고, 단어음에 대한 인식실험을 수행하였다.

인식실험에 사용할 단어음 데이터는 총 6500개의 고립 단어음이다. 이들 음성 데이터는 남성화자 10인으로부터 각 50개의 단어음에 대해 각 13회씩 발성한 음성을 채집한 것이다. 그리고 음소분류신경망에 사용할 음소 데이터는 인식에 사용할 단어음에 내포된 음소 23개에 대해 20회씩 발음한 4600개로 구성하였다. 이들 단어음과 음소데이터는 인식시스템의 학습데이터와 인식실험에 사용할 데이터로 나누어 사용된다. 각 음성데이터는 선형예측 분석방법을 통하여 얻은 LPC 계수로부터 16차 LPC 켑스트럼 계수와 로그에너지 및 평균 영교차율로 구성된 18개의 요소를 갖는 벡터를 추출하여 그 음성 프래임의 특징벡터로 이용하였다. 이렇게 얻어진 음성의 특징벡터는 음소분류 신경망의 입력으로 사용하였다. 이때 음성의 시간지연 특성을 고려하기 위하여 인식실험에서는 현재(t)의 음성프래임과 이전(t-1) 프래임을 이용하여 음소분류신경망의 입력으로 사용하였다. 따라서 음소분류신경망의 입력층의 노드 수는 36개가 된다.

구현한 인식시스템에서 음소분류신경망의 성능은 최종적으로 단어음 인식에서 결정적인 역할을 하게된다. 따라서 음소분류인식기의 성능을 확인하기 위해 음소인식 실험을 수행하였다. 이를 위해 음소 데이터 중에서 2300개는 음소분류신경망의 학습에 이용하고, 나머지는 2300개는 인식실험에 사용하였다. 음소분류신경망의 각 음소그룹별 다층신경망의 학습은 모멘텀률 α를 0.7로 하여 학습 오차값이 0.001이하가 되도록 학습하였다. 음소인식 인식실험에 대한 오인식 결과는 음소특성에 따라 음소를 그룹별로 분류하지 않은 경우 17.5%의 오인식률 보였다. 그리고 특성별로 분류한 경우에서는 14.4%의 오인식 결과를 나타냈으며, 분류하지 않은 경우보다 전체 오인식률이 3.1% 향상된 음소인식 결과를 보였다. 그리고 적응인식 결과는 10.7% 였다. 음소분류신경망에 의한 음소인

식 실험결과는 전체적으로 비교적 음성특징이 잘 구별되는 모음에서는 높은 인식률을 보였지만 자음에서는 유사 특성을 갖는 음소간의 혼동으로 인식률이 저조함을 보였다. 그리고 단어음 인식실험을 위해 50개의 단어음을 10명이 13회씩 발성한 6500개의 단어음 데이터중에서 5명이 50개의 단어음에 대해 13회씩 발성한 3250개의 단어음에 대해서는 HMM의 학습데이터로 이용하였다. 이 50개의 단어음에는 한 개의 음절로 이루어진 숫자음과 둘에서 세 음절로 이루어진 지명을 포함하고 있다. 먼저, HMM의 학습에 참여한 5명의 대해 화자종속 인식실험을 수행해 보았다. 이를 표 1에서 각 화자들에 대해 각 음절별로 오인식률을 보이고 있다. 실험결과는 전체 평균 5.9%의 오인식률을 보였다.

표 1. 화자종속 음성인식에서 각 화자별 오인식률(%)
Table 1. Error rates(%) of each speaker in speaker-dependent speech recognition.

구분 / 화자	1음절	2음절	3음절	평균
A	6.3	1.4	1.5	3.1
B	19.5	6.3	6.1	10.6
C	6.3	1.6	3.1	3.7
D	8.4	1.8	1.5	3.9
E	16.1	4.5	4.6	8.4

그리고 HMM 학습에 참여하지 않은 나머지 5명이 50개의 단어음을 13회 발성한 3250개의 단어음 중에서 3회씩 발성한 750개의 단어음을 인식시스템을 적응시키기 위한 교정데이터로 사용하고, 나머지 5회 발성한 2500개의 단어음을 인식실험 데이터로 사용하였다. 이 인식실험 데이터를 이용하여 화자독립 인식실험을 수행하여 그 결과를 표 2에 보였다. 실험결과, 전체 평균 15.4%의 오인식 결과를 보였으며, 화자적응을 수행하지 않은 상태이기 때문에 화자종속의 평균 오인식에 비해 상당히 인식률이 떨어짐을 알 수 있다. 표 3에는 HMM 학습에 참여하지 않은 5명의 화자에 대해 3개의 교정데이터를 이용하여 표준화자의 음성 특징 파라미터를 새로운 화자의 음성 특징 파라미터에 적응시켜 인식한 실험결과를 보였다. 화자적응시 전체 평균 오인식률은 9.0%로서, 화자독립 음성인식에서 저조한 인식성능을 1차선형 변환망에 의해 화자적응을 시켜 인식한 결과, 화자독립에 비해 전체 평균 오인률을 6.4% 감소시킬 수 있었다. 본 화자적응 인식시스템은 화자적응을 위해 음소분류신경망의 전단의 변환망의 파라미터만을 조정하여 새로운 화자에 적응시키고 적은 수의 교정데이터로도 만족할 만한 인식성능을 얻을 수 있다. 그리고 음소분류신경망도 우리말 음소특징에 따라 그룹화하여 각각의 신경망을 구성하여 음소변별력을 증가시켜 HMM 음성인식 성능향상에 도움이 되도록 하였다. 그리고 음소분류신경망은 모든 HMM에서 공통적으로 이용하게 하였다.

표 2. 화자독립 음성인식에서 각 화자별 오인식률(%)
Table 2. Error rates(%) of each speaker in speaker-independent speech recognition.

구분 / 화자	1음절	2음절	3음절	평균
F	20.0	5.9	7.6	11.2
G	20.0	3.5	6.2	9.9
H	21.8	18.8	21.5	20.7
I	17.3	20.9	20.0	19.4
J	15.5	16.2	15.4	15.7

표 3. 화자적응 음성인식에서 각 화자별 오인식률(%)
Table 3. Error rates(%) of each speaker in speaker-adaptation speech recognition.

구분 / 화자	1음절	2음절	3음절	평균
F	14.8	1.5	3.1	6.5
G	16.4	1.5	6.2	8.0
H	10.0	8.8	6.2	8.3
I	7.3	12.3	12.3	10.6
J	9.1	15.0	10.8	11.6

V. 결론

본 논문에서는 화자독립 고립 단어음을 인식하기 위한 화자적응 음성인식 시스템을 구성하여 실험하였다. 인식시스템은 그 전단에 음소분류신경망을 이용한 반연속 HMM을 기반으로 하고 있다. 음소분류신경망은 음소특징별로 그룹화하여 변별력 높이도록 하여 HMM단어음 인식률을 개선할 수 있게 하였다. 각 음소그룹은 다층퍼

셉트론으로 각각 구성하였다. 그리고 화자적응 인식을 위해 음소분류신경망의 전단에 변환망을 두고, 인식시에는 표준화자의 특징벡터를 새로운 화자의 음성특징에 선형적으로 적응하여 인식하게 하였다. 그 결과, 구현한 화자적응 인식시스템은 고립 단어음 인식실험에서 오인식률이 화자독립 인식의 15.4%에 비하여 화자적응 인식이 9.0%로 감소되었다.

참고문헌

[1] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," IEEE Acoust., Speech, Signal Processing, Mag., pp. 4-16, Jan. 1986.

[2] B. H. Juang, "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Obsevations of Markov Chains," AT&T Tech J., Vol. 64, pp.1235-1249, July-Aug. 1985.

[3] X. D. Huang, Y. Ariki, and M. A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, England, 1990.

[4] B. F. Necioglu, M.Ostendorf, and J. R. Rohlicek, "A Bayesian Approach to Speaker Adaptation for the Stochastic Segment Model," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 449-452, 1992.

[5] H. Matsukoto and H. Inoue, "A Piecewise Linear Spectral Mapping for Supervised Speaker Adaptation," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 437-440, 1992.

[6] O. Schmidbauer and J. Tebelskis, "An LVQ Based Reference Model for Speaker-Adaptive Speech Recognition," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.441-444, 1991.

[7] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. on Acoust., Speech, Signal Processing, Vol. ASSP-26, pp. 43-49, Feb. 1978.

[8] P. L. Cerf, W. MA and D. V. Compernolle, "Multilayer Perceptrons as Labelers for Hidden Markov Models," IEEE Trans. on Speech and Audio Processing, Vol. 2, No.1, Part II, pp. 185-193, Jan. 1994.

[9] H. Bourland and C. J. Wellekens, "Links between Markov Models and Multi-layer Perceptrons," Advanced in Neural Information Processing Systems, Vol. 1, pp. 502-510, 1989.

저 자 소 개

이 기 희

제 2 권 제 2 호 참조

현재 동서울대학 사무자동화과 조교수