인지 모델을 이용한 제한된 한국어 연속음 인식

Recognition of Restricted Continuous Korean Speech Using Perceptual Model.

김 선 일*, 홍 기 원**, 이 행 세**
(Seonil Kim*, Kiwon Hong**, Haing Sei Lee**)

요 약

본 논문에서는 사람의 인지 특성에 가까운 PLP 캡스트럼을 사용하여 음성의 시간적 특성을 잘 반영할 수 있도록 넓은 시간대에 걸쳐 특징을 추출하였으며 인간의 학습 방법과 유사한 인공신경망을 이용하여 음소를 인식하고 인식된 음소로부터 순서 특징을 잘 반영하는 Markov 모델을 통해 음소열을 인식하였다.

음소인식은 연속음성에 나타나는 음소에서 비균일한 프레임 갯수로 채취된 음성 블럭들을 사용하여 7차 PLP 캡스트럼, PTP, 영교차율 및 에너지를 구하고 이를 MLP 신경망의 입력으로 사용하여 두 사람이 각각 5번씩 발음한 10 종류의 한국어 문장, 총 100개를 대상으로 음소 인식을 실시하여 최대 94.4%의 음소별 인식률을 얻을 수 있었다.

문장인식은 학습에 참여했던 두 사람이 각 문장에 대해 10번씩 새로 발음한 총 200개의 데이타에 대해 음소별 인식을 거쳐 첫번째 실험을 통해 생성된 Markov 모델을 이용하여 문장 인식을 실시한 결과 92.5%의 문장 인식률을 얻었다.

Abstract

In this paper, the PLP cepstrum which is close to human perceptual characteristics was extracted through the spread time area to get the temperal feature. Phonemes were recognized by artificial neural network similar to the learning method of human. The phoneme strings were matched by Markov models which well suited for sequence.

Phoneme recognition for the continuous Korean speech had been done using speech blocks in which speech frames were gathered with unequal numbers. We parameterized the blocks using 7th order PLPs, PTP, zero crossing rate and energy, which neural network used as inputs. The 100 data composed of 10 Korean sentences which were taken from the speech two men pronounced five times for each sentence were used for the the recognition. As a result, maximum recognition rate of 94.4% was obtained.

The sentence was recognized using Markov models generated by the phoneme strings recognized from earlier results the recognition for the 200 data which two men sounded 10 times for each sentence had been carried out. The sentence recognition rate of 92.5% was obtained.

*거재전문대 전자과
(Dept. Electronics, Geoje Junior College)
**아주대학교 전자공학과
(Dept. of Elec. Eng., Ajou Univ.)
접수일자 : 1995년 1월 25일

Ⅰ. 서 론

사람이 의사를 전달하는 가장 기본적인 도구는 음성이다. 기록으로 남기기 위한 의사 소통 도구로서 문자가 존재하지만 가장 쉽게 가장 자주 사용되는 것이 음성을 통한 정보의 교환이다. 최근에는 인간과 인간 사이의 정보 교환외에 인간과 기계 사이의 의사 소통도 중요한 상황이 되었으며 이를 위한 노력이 꾸준히 이루어져 왔다. 인간과 기계 사이의 정보 교환에는 여러가지 도구들이 쓰이고 있지만 그 중 가장 자연스러운 것이 음성이라 할 것이다. 기계에 음성을 인식시키기 위해 여러가지 방법들이 개발되어 왔고 해당 음성을 가장 잘 나타내는 특징들을 찾기 위해 부단한 노력이 이루어져 왔다. 그러나 기계가 인간과 같은 능력으로 인식하기에는 아직도 갈 길이 먼 것처럼 보인다. 따라서 대상 어휘를 제한하여 고립 단어나[11] 숫자의 인식[12]등 제한 영역에서 응용하고자 하는 노력들이 이루어져 왔다. 응용의 폭을 넓히려는 경우 대규모의 데이타베이스를 필요로하며 많은 연산량이 요구된다[7].

인간의 청각은 주파수의 변화에 따라 소리를 인지하며 각 주파수 영역에서의 인지도가 서로 다르다. 기계에 음성을 인식시키고자 할 때 사람과 유사한 형태로 시도하는 것이 가장 바람직한데 음성 특징을 추출할 때는 사람의 청각 신경을 모방한 인지선형예측법(PLP : Perceptual Linear Prediction)[8, 9, 10]을 사용하여 화자 독립적인 음성의 특징을 구하고 각 음소를 인식할 때는 인공 신경망을 사용하여 훈련된 음소에 대해 인식을 수행하였다.

간단한 문장은 사람들 사이에 자주 쓰이면서 언어를 이용한 의사 소통에 자주 사용된다. 문장의 완성도에 비해 그 사용빈도는 극히 높은 편이다. 따라서 자주 쓰이는 10개의 문장을 대상으로 문장 인식을 실시하였다.

사람들은 불완전한 음성이 들어오더라도 자신이 가진 데이타베이스를 이용하여 가장 가까운 문장으로 사상시키는 능력이 있다. 이런 기능은 Markov model을 이용하여 구현하였다. 신경망으로 출력된 음소열을 이용하여 각 문장의 Markov 모델을 구하고 음소 인식 신경망에서 나오는 음소열을 Markov 모델에 적용시켜 최대의 값을 내는 모델을 인식된 것으로 인정하였다.

본문에서는 먼저 인지선형예측 분석법과 Markov 모델의 이론적 근거를 살펴보았다. 그 후에 인식시스템에 대한 설명을 하고 PLP 등을 특징값으로 사용하고 비균일한 블럭으로 구성되는 입력 벡터를 이용한 MLP 신경망에 의한 음소 인식 실험 결과와 Markov 모델에 의한 문장 인식 실험 결과를 제시하였다.

Ⅱ. 인지선형예측 분석법[8]

PLP모델은 더 낮은 차수의 전극점 모델의 스펙트럼에 의해 근사된다. 귀가 느끼는 스펙트럼은 0-5kHz의 주파수 범위에서 16개의 대역들로 나뉘어서 합하여지며, 중간대역과 상위대역을 보강하기 위하여 equal-loudness pre-emphasis를 거치게 된다[13]. 또한 음성 스펙트럼의 전력 변화율을 감소시키기 위하여 3제곱근 처리를 함으로써 intensity-loudness 3 제곱근 크기 압축을 실시한다[3]. 이러한 처리를 거친 16개의 스펙트럼 성분에 푸리에 역변환 과정을 적용시켜 자기 상관 계수[1, 2]를 얻는다. 전극점 모델은 얻어진 자기상관 계수로부터 원하는 차수로 계산되며, 이로부터 다시 켑스트럼 계수를 계산할 수 있다. 본 논문에서는 7 차의 전극점 모델을 사용하였다.

음성신호의 세그먼트는 해밍창(Hamming Window)에 의해 처리된다.

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

단, N은 창의 길이이다. FFT를 위하여 25. 6ms(256 point)의 시간창을 사용하였으며, 단구간 제곱합을 구하기 위하여 전력 스펙트럼은 식(2)와 같이 실수성

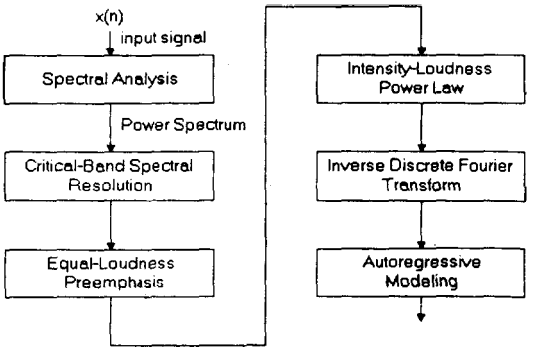

그림 1. PLP 분석법의 블럭도
Fig 1. Block diagram of PLP speech analysis method

분과 허수성분을 제곱하여 더한다.

$$P(\omega)=\mathrm{Re}[S(\omega)]^2+\mathrm{Im}[S(\omega)]^2 \quad (2)$$

스펙트럼은 다음과 같은 Bark frequency Ω에 의해 주파수 축을 따라 굴절된다.

$$\Omega(\omega)=6\ln\left\{\frac{\omega}{1200\pi}+\left[\left(\frac{\omega}{1200\pi}\right)^2+1\right]^{0.5}\right\} \quad (3)$$

단, ω는 rad/s의 각속도이다. 이 Bark-Hertz 변환은 Schroeder(1977)[13]에 의해 제안되었다. 결과치인 굴절된 전력스펙트럼은 임계대역 마스킹 곡선(critical-band masking curve) $\Psi(\Omega)$와 콘볼루션된다. 임계대역곡선(critical-band curve)은 다음과 같이 주어진다.

$$\Psi(\Omega)=\begin{cases}0 & \text{for } \Omega<-1.3\\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3\le\Omega\le-0.5\\ 1 & \text{for } -0.5<\Omega<0.5\\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5\le\Omega\le2.5\\ 0 & \text{for } \Omega>2.5\end{cases} \quad (4)$$

$\Psi(\Omega)$과 $P(\Omega)$의 이산 콘볼루션은 임계대역 전력스펙트럼 $\Theta(\Omega_i)$를 만든다.

$$\Theta(\Omega_i)=\sum_{\Omega=-1.3}^{2.5}P(\Omega-\Omega_i)\Psi(\Omega) \quad (5)$$

상대적으로 넓은 대역의 임계대역 마스킹 곡선 $\Psi(\Omega)$와의 콘볼루션은 원래의 $P(\Omega)$에 비하여 스펙트럼 분해능이 심하게 저하된다. 제안한 방법에서는 대략 1-Bark 간격으로 $\Theta(\Omega)$를 표본화 하였다. 정확한 수의 표본화 간격은 주파수의 표본들이 분석하는 대역의 모든 주파수 표본들을 포함하여야 한다. 보통은 18개의 표본치들이 0~5 kHz영역을 포함하도록 하지만 본 논문에서는 FFT의 편의를 위하여 16개의 표본을 취하였다.

그림 2는 표본화된 여과기들을 나타낸다. 16개의 여과기들이 0Hz에서 5kHz를 포함하도록 되어있으며, 가로축은 주파수의 정상적인 스케일을 나타내며 세로축은 각 필터의 주파수에 대한 감쇄비율을 나타낸다. 즉 세로축이 $\Psi(\Omega)$를 나타내며 가로축은 주파수를 나타낸다.

표본화된 $\Theta(\Omega)$는 근사된 equal-loudness 곡선에

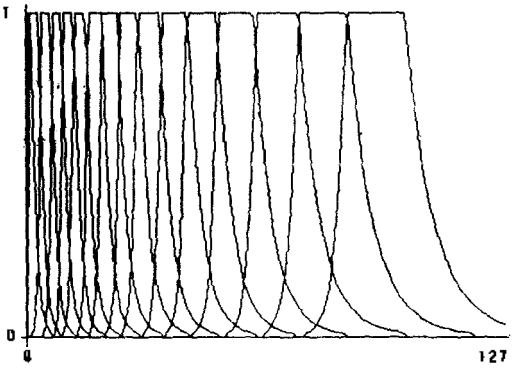

그림 2. 임계대역주파수해법에 의한 여과기
Fig 2. The filters from critical-band spectral resolution

의해 여과된다.

$$\Xi[\Omega(\omega)]=E(\omega)\,\Theta[\Omega(\omega)] \quad (6)$$

$E(\omega)$는 서로 다른 주파수들에 대하여 사람이 다른 민감도를 갖는것을 근사하고, 약 40dB의 정취 감도를 모방한다 이것에 대한 근사함수는 다음과 같다.

$$E(\omega)=\frac{(\omega^2+56.8\times10^6)\omega^4}{(\omega^2+6.3\times10^6)^2(\omega^2+0.38\times10^9)} \quad (7)$$

그림 3은 이 근사함수를 도식적으로 나타낸 것으로서 가로축은 주파수의 정상 스케일을 나타내며 세로축은 필터의 응답을 나타낸것이다.

식 (7)은 0과 400Hz 사이에서 12dB/oct, 400과

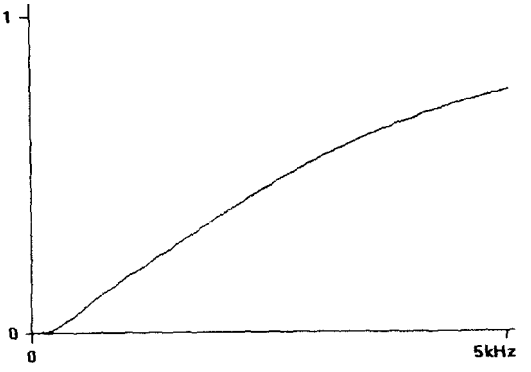

그림 3. Equal-loudness pre-emphsis 여과기의 특성
Fig 3. The characteristics of equal-loudness pre-emphasis filter

1200Hz 사이에서 9dB/oct, 1200과 3100Hz 사이에서 6dB/oct 그리고 3100Hz 이상에서 0dB/oct의 감쇄율을 가지는 선달함수를 표현한다. 보통의 소리 수준에서는 이러한 근사방법이 5000Hz까지 매우 정확하다. 하지만 더 높은 대역폭을 갖는 응용에 있어서는 5000Hz 이상에서 더 급격한 감쇄율을 갖는 항을 첨가해야한다.

전극점 모형화 전의 처리단계중의 마지막은 3 제곱근 크기 압축(cubic-root amplitude compression)이다.

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \tag{8}$$

이 처리 단계는 청각의 진력법칙을 근사하고 소리의 세기와 귀가 느끼는 세기의 비선형성을 모방한다. 또한 전극점 모델이 더 적은 차수의 계수를 가지도록 스펙트럼의 변화를 줄이는 작용을 한다.

처리과정을 거친 스펙트럼은 결과적으로 전력 스펙트럼이므로 푸리에 역변환 과정을 거쳐 자기상관 함수를 얻을 수 있다. 임계대역 분석을 거치므로써 주파수 분해능이 현저히 낮아져 있으므로 낮은 차수의 역변환으로 원하는 차수의 자기상관 계수를 얻을 수 있다. 본 논문에서는 16개의 Bark-spectrum을 사용하였으므로 32-point FFT를 적용하였다. 이와 같이 얻어진 자기상관 계수들은 전극점 모델을 구하는 자기회귀처리에 직접 이용할 수 있으며 또한 켑스트럼 계수와 같이 다른 계수를 구하는데 이용할 수 있다.

PLP분석을 위한 연산 복잡도는 LP분석에 비해 대단히 크다. 연산측면에서 가장 복잡한 부분이 FFT 스펙트럼 계산과, 이어지는 임계대역 스펙트럼 적분과 3 제곱근 압축등이다. AR 모델을 위한 연산의 복잡도는 주파수 분해능이 낮아져서 서로 상쇄될 수 있다.

전극점 모델의 결과 스펙트럼은 일반 LP 모델의 스펙트럼에 비해 더 선형적이다[10]. 또한 일반 LP 모델에 비해 더 낮은 차수의 모델링이 가능하다[10]. 결과적으로 인공신경망의 입력의 감소와 데이타베이스의 역할을 하는 가중치들을 줄이는 역할을 하게되며, 이것은 처리속도의 효율화에 기여하게 된다[5].

Ⅲ. Markov 모델

신경망은 분류 기능이 강력한데 비해 Markov 모델은 시간적 특성을 잘 구분해내는 stochastic temperal network 으로서 신경망으로 인식된 결과로 얻게되는 음소열로부터 그 음소열이 어떤 문장에 해당되는지를 알아내는 후처리 방법으로 사용할 수 있다. 일반적으로 음성 인식에 많이 사용하는 HMM(Hidden Markov Model)은 이중 통계 모델(doubly stochastic model)인데 반해 Markov 모델은 단일 통계 모델(singly stochastic model)로서 관측 상태가 바로 출력값으로 관찰이 되기 때문에 구현이 쉽고 재평가(reestimation) 문제를 통한 최적화 문제가 없기 때문에 구태여 HMM 처럼 방대한 통계 처리가 필요없으므로 데이타 수집상의 문제를 해결할 수 있다. 그리고 신경망과 Markov 모델의 결합은 분류기능과 음성 처리에 꼭 필요한 시간 특성 포착 기능을 같이 처리할 수 있다는 점에서 상당한 잇점을 가지고 있다.

L 개의 모델이 있다고 가정할 때 각 모델은 $=\langle A, \pi\rangle$로 모델링 되는데 여기에 쓰이는 변수들은 다음과 같다.

$Q=\{q_1, q_1,...,q_N\}$ 는 N 개의 있을 수 있는 상태를 나타낸다.

$\pi=\{\pi_1, \pi_2,...,\pi_N\}$, 그런데 $\pi_i = Pr(q_i \text{ at } t=1)$ 는 초기상태 분포확률값이다.

A = 상태 천이 행렬 $\{a_{ij}\}$, 그런데 $a_{ij} = Pr(q_j \text{ at } t+1 \mid q_i \text{ at } t)$ 는 상태 천이 확률이다.

$O=\{O(1),...,O(T)\}$, 전체 길이가 T 인 관측열로서 O(t)는 $Q=\{q_1, q_2,...,q_N\}$ 에 있는 상태중의 하나이다.

Markov 모델의 학습은 통계적 계산에 의해

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}}{\sum_{t=1}^{T-1} \gamma_i} \tag{9}$$

$\sum_{t=1}^{T-1} \gamma_i(t)$ = 상태 q_i로부터 상태 천이되는 총횟수

$\sum_{t=1}^{T-1} \xi_{ij}(t)$ = 상태 q_i에서 상태 q_j로 천이가 일어나는 총횟수

로 구성된 상태 천이 행렬과 π_i를 구하면 된다.

Ⅳ. 인식시스템

인식 시스템은 10 kHz 샘플링 주파수를 갖는 12 bits A/D, D/A 변환기를 갖는 컴퓨터 및 소프트웨어로 작성된 MLP 신경망과 후처리를 위한 Markov 모델로 구성된다. MLP 신경망으로는 하나의 은닉층을 갖는 오차역전달 신경망을 사용하였으며 출력층은

19개의 노드(node) 로 구성되고 은닉층은 40개의 노드로 구성하였다. 신경망 입력으로는 3 ms 마다 256 표본을 취하여 계산된 7차 PLP[8, 9] 캡스트럼(cepstrum) 계수[12]와 영교차율, 단구간 에너지 및 단구간 PTP(Peak To Peak) 값을 사용하였다. PLP와 영교차율, 에너지를 특징값으로 사용하면 입력 벡터 하나당 9개의 특징값을 갖게 되고 따라서 입력 벡터를 7개 사용하면 입력 노드는 63개가 되고 입력 벡터를 5개 사용했을 경우 45개의 입력 노드를 갖게 된다. 특징값으로 단구간 PTP를 추가하면 벡터 하나당 10개의 특징값을 갖게 되므로 7개의 입력 벡터를 사용할 경우 70개, 5개의 입력 벡터를 사용할 경우 50개의 노드를 갖게 된다. 출력함수로는 시그모이드(sigmoid) 함수를 사용하였다. PLP를 이용한 캡스트럼 계수는 PLP 계수로부터 다음과 같이 얻을 수 있다.

$$C_i = a_i + \sum_{j=1}^{i-1} C_j \cdot a_{i-j} \qquad i = 1, 2, \cdots, p \tag{10}$$

PLP 계수를 포함한 LP계수는 차수가 올라갈 수록 작은 크기의 값을 갖는다. 본 논문에서는 모든 계수가 비슷한 범위를 갖도록 다음 차수의 값을 곱하여 표현하였다. 즉,

$$C_i = C_i \cdot (i+1) \qquad i = 1, 2, \cdots, p \tag{11}$$

학습패턴은 인식기의 학습을 위한 목표 값이며 올바른 목표치를 사용하여야 올바르게 학습을 시킬 수 있다[4, 5]. 음성은 연속적인 성질을 가지고 있으며 조음현상의 영향으로 인하여 음의 경계가 분명치 않다. 또한 비슷한 입력에 대해 다른 출력을 요구하는 자체도 인공신경망의 학습시간을 연장시키며 또한 잘못된 입력공간을 구성해 목표치에 수렴하지 못할 가능성이 높다[4, 5]. 그러므로 그림 4 같이 음소의 경계 부분 즉 전이구간에 해당하는 부분은 학습 목표 구간에서 제외시키고 그 외의 부분을 목표치로 설정하면 학습의 혼돈이 방지될 수 있다. 이와 같이 하지 않으면 음소경계에서는 애매한 출력, 즉 경계 양쪽 음소의 특징을 모두 포함하는 출력을 내도록 학습될 것이므로 전체 학습 및 인식에 상당한 지장을 초래하게 된다.

그림 4 는 목표치를 설정하는 방법을 나타낸 것이다. 전체를 한 문장에 대한 음성 파형이라고 가정했을 때 검고 두껍게 나타낸 부분을 주위 음소의 학습에 영향을 미치지만 자신은 학습되지 않는 부분으로 설정하였으며 실험에 사용된 음성을 분석한 결과 평균 20ms로 설정하였다.

인공신경망의 학습 및 인식을 위하여 인공신경망의 입력단 및 학습 벡터의 구성을 그림 5와 같이 하였다. 네모칸 안에 쓰인 숫자는 구간의 지속시간이며 밑의 숫자는 현재 위치로부터 전과 후의 시간이다.

연속음에 나타나는 음소는 전후의 다른 음소와의 조음 현상에 의한 변화가 심하다. 현재의 음성 조각을 인식하기 위해서는 인접한 음소를 같이 참조하여야 자연스럽다. 보통은 이러한 변화를 학습하기 위하여 시간지연신경망을 사용하지만 본 논문에서는 고정 신경망을 사용하고, 입력단의 구성을 시간지연신경망과 다른 형태로 시간적 특성을 수용하였다.

현재 프레임 3ms에 대해 그 전후까지 고려한 특징들의 집합을 구성하는데 그림 5 와 같이 부등간격으로 취해진 각 블럭(block)을 하나의 벡터로 생각하면 한 벡터가 7차 PLP 캡스트럼계수들과 단구간 에너

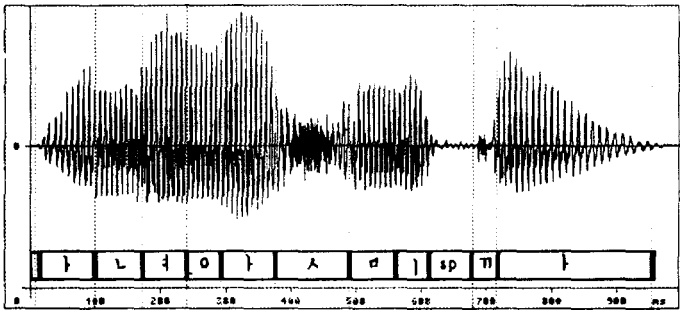

그림 4. 학습을 위한 목표치의 설정
Fig 4. The setting of teaching vector for learning

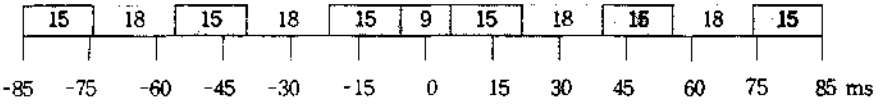

그림 5. 입력벡터의 요소 구성
Fig 5. The distribution of 70 PLP features

지 및 단구간 영교차율, 단구간 PTP, 총 10개의 요소로 구성되었을 경우 7개의 벡터를 사용하면 전체 70개의 데이타가 신경망의 입력으로 구성된다.

네모칸 안의 숫자는 각 블럭의 시간폭을 나타낸다. 실선은 PLP가 평균되어지는 구간의 길이를 나타내며, 점선은 사용되지 않는 구간의 길이이다. 9는 현재의 프레임(3ms) 및 전 프레임(3ms), 후 프레임(3ms), 총 3개의 프레임으로 한 블럭을 형성한 것을 나타내며 세 프레임에서 구한 특징들이 평균되어 한 특징 벡터로 나타내어진다. 15는 5개의 프레임(총15ms)으로 한 블럭을 형성한 것을 나타내며 다섯 프레임에서 구한 특징들이 평균되어 한 특징 벡터로 나타내어진다. 18은 사용되지 않는 구간인데 6개의 프레임 즉 18ms 길이이다.

사용된 자모는 묵음까지 19개이며 19개의 상태를 갖는 Markov 모델로 모델링하여 후처리 하였다.

V. 실험 및 검토

실험은 크게 두가지로 나누어서 수행했는데 먼저, 신경망을 이용하여 사용된 특징값을 달리하고 입력 벡터의 수를 변화시켰을 때 각 문장내의 음소의 인식률에 대해 조사하고 다음으로 신경망 출력에 대해 Markov 모델을 구성하여 문장 인식을 실시하였다.

1. 음소 인식 실험

실험 대상 음성으로는 먼저 두 사람이 10 종류의 한국어 문장 "안녕하십니까", "감사합니다", "고맙습니다", "계속합니다", "그만합니다", "어서오세요", "믿겠습니다", "시작합니다", "사랑합니다", "죄송합니다"를 한 사람당 다섯번씩 발음하게 하여 총 100개의 문장을 수록한 데이타를 사용하였다. 다섯번 발음한 것 중 각 사람당 한 문장씩 학습에 참여하게 하여 각 사람 당 10문장 총 20개의 문장을 학습에 사용하였고, 나머지 80 개의 문장과 학습에 참여한 20 문장을 합쳐 100개의 문장에 대해 인식 실험을 실시하였다. 문장의 프레임별 인식 결과가 9ms 이내로 나타나는 것은 신뢰성이 작으므로 여과시켜 제거하고 충분히 긴 것에 대해서 결과를 얻었다.

PLP 켑스트럼과 영교차율, 에너지를 사용하여 입력 벡터가 7개일 때와 5개일 때 화자 A와 화자B에 대해서 인식 결과를 표 1에 나타내었다. 표 1의 (a)에는 입력 백터가 7개일 때 음소 인식 결과를 나타내었다. 표 1의 (b)는 입력 벡터가 5개일 때 음소 인식 실험 결과이다. 표의 결과에서 '/' 뒤에 나타나는 숫자는 총 음소 수이고 '/' 앞에 나타나는 수는 인식된 음소 수이다. 예를 들어 "안녕하십니까"는 "ㅏㄴㅕㅇㅏㅅㅁㅣspㄲㅏ"의 음소로 구성되는데(여기서 하 은 아 에 흡수되고, '시'의 'ㅣ'는 'ㅅ'에 흡수되어 거의 학습이 되지 않아 제외시켰음) 총 11 개의 음소로 구성되고 한 사람이 한 문장을 다섯번 발음하였으므로 전체 음소 수가 55개 이다. 전체 문장의 음소 수를 합하면 'ㄱ'의 경우 네 문장에서 한번식 나타나므로 다섯번 발음하면 총 20 개가 나타나고 화자 B의 경우 그 중 16개가 인식되었다. 인식률을 보면 5개의 백터를 사용한 경우(90.7%)보다 근소한 차나마 7개를 사용한 경우(91.8%)가 더 높은 인식률을 보여주고 있다.

"죄송합니다"의 경우는 복모음이 들어가 있어서 "제송합니다"로 발음하였다. (실제로 사람들이 무심코 발음할 때면 이렇게 발음하는 경우가 많다.) 인식 결과를 보면 몇몇 자음에 대해 인식률이 저조한데 자음 인식의 어려움을 단적으로 증명해 주고 있다.

표 2는 음성 특징으로 PTP를 추가해서 실험한 인식 결과로서 표 2의 (a)는 입력 벡터가 7개 일 때 음소 인식 결과를 보여주고 있으며 표 2의 (b)는 입력 벡터가 5개일 때 각 음소 인식 결과를 보여주고 있다. 표 2를 보면 입력 벡터가 5개일 때의 인식률(89.9%)보다 7개일 때의 인식률(94.4%)이 훨씬 좋게 나타나 있다.

인식 결과를 보면 화자에 따라 인식률에 약간의 차가 있음을 알 수 있다. 연속 음성을 발음해 보면 알아듣기 쉬운 음성이 있고 그렇지 않은 음성이 있어서

표 1. PLP 캡스트럼, 영교차율, 에너지를 특징값으로 사용했을 때의 음소 인식 결과
a) 입력 벡터가 7개일 때 각 음소별 인식 결과
b) 입력 벡터가 5개일 때 각 음소별 인식 결과

Table 1. Phoneme recognition results using energy, zero crossing rate and PLP cepstrum as features
a) Phoneme recognition results per phoneme with 7 input vectors
b) Phoneme recognition results per phoneme with 5 input vectors

음 소	화자 A	화자 B	인 식 율
ㄱ	6/20	16/20	55.0%
ㄲ	5/10	2/10	35.0%
ㄴ	7/10	4/10	55.0%
ㄷ	40/40	40/40	100%
ㅁ	60/65	58/65	90.8%
ㅅ	40/40	39/40	98.8%
ㅆ	10/10	10/10	100%
ㅇ	13/15	14/15	90.0%
ㅈ	6/10	7/10	65.0%
ㅋ	4/10	10/10	70/0%
ㅏ	108/110	108/110	98.2%
ㅓ	10/10	10/10	100%
ㅕ	5/5	5/5	100%
ㅗ	15/20	20/20	87.5%
ㅛ	4/5	5/5	90.0%
ㅡ	2/5	5/5	70.0%
ㅣ	55/55	55/55	100%
ㅐ	20/20	18/20	95.0%
sp	20/20	20/20	100%
TOTAL	430/480 89.6%	451/480 94.0%	881/960 91.8%

(a)

음 소	화자 A	화자 B	인 식 율
ㄱ	17/20	18/20	87.5%
ㄲ	5/10	3/10	40.0%
ㄴ	7/10	9/10	80.0%
ㄷ	39/40	40/40	98.8%
ㅁ	58/65	55/65	86.9%
ㅅ	38/40	38/40	95.0%
ㅆ	10/10	10/10	100%
ㅇ	11/15	11/15	73.3%
ㅈ	4/10	2/10	30.0%
ㅋ	5/10	7/10	60.0%
ㅏ	109/110	109/110	99.1%
ㅓ	6/10	8/10	70.0%
ㅕ	5/5	4/5	90.0%
ㅗ	20/20	19/20	97.5%
ㅛ	5/5	5/5	100%
ㅡ	4/5	4/5	80.0%
ㅣ	55/55	53/55	98.2%
ㅐ	20/20	18/20	95.0%
sp	20/20	20/20	100%
TOTAL	438/480 91.3%	433/480 90.2%	871/960 90.7%

(b)

각자 개인차를 뚜렷이 가지고 있다. 이를 비교해보려면 표준 음성 데이타 베이스가 있어야 할 것으로 생각된다. 전체적으로 보면 에너지, 영교차율, PLP, PTP를 특징값으로 사용하고 입력 벡터가 7개일 때 좋은 결과를 나타내고 자음보다는 모음에서 좋은 인식률을 나타내고 있다.

학습과 인식에 사용된 음소는 각 문장에서 추출한 것으로 문장에서 나타나는 빈도에 따라 사용된 갯수가 다른데 총 100 개의 문장 중 'ㅏ'가 가 제일 많은 220회나 나타나고 'ㅕ', 'ㅛ', 'ㅡ'가 가장 적어 10회 나타난다.

2. 문장 인식 실험

인식된 음소로부터 문장을 인식하기 위한 후처리 과정으로서 N=19, 즉 19개의 상태를 갖는 Markov 모델을 사용하였다. 음소 인식에는 앞에서 나온 실험 결과에 따라 가장 좋은 결과를 보이는 것을 선택하였는데 음성 특징으로 PLP, 영교차율, 에너지, PTP를 사용하고 7개의 벡터를 구성하여 신경망의 입력으로 사용하였다. 따라서 입력 노드 70개, 은닉 노드 40개, 출력 노드 19개로 신경망을 구성하였다. Markov 모델은 관찰열 자체가 상태열이므로 구현이 간단하다. 10개의 문장을 두 사람에게 각각 5번씩 발음하게 하여 각 사람에게서 50개의 데이타 즉, 총 100개의 데이

표 2. PLP 켑스트럼, PTP, 영교차율, 에너지를 특징값으로 사용했을 때의 음소 인식 결과
a) 입력 벡터가 7개일 때 각 음소별 인식 결과
b) 입력 벡터가 5개일 때 각 음소별 인식 결과

Table 2. Phoneme recognition results using energy, zero crossing rate PTP and PLP cepstrum as features
a) Phoneme recognition results per phoneme with 7 input vectors
b) Phoneme recognition results per phoneme with 5 input vectors

음 소	화자 A	화자 B	인 식 율
ㄱ	17/20	19/20	90.9%
ㄲ	8/10	7/10	75.0%
ㄴ	4/10	10/10	70.0%
ㄷ	40/40	40/40	100%
ㅁ	65/65	60/65	96.2%
ㅅ	39/40	38/40	96.3%
ㅆ	10/10	10/10	100%
ㅇ	12/15	12/15	80.0%
ㅈ	5/10	3/10	40.0%
ㅋ	8/10	9/10	85.0%
ㅏ	110/110	109/110	99.5%
ㅓ	10/10	10/10	100%
ㅕ	5/5	4/5	90.0%
ㅗ	19/20	19/20	95.0%
ㅛ	4/5	5/5	90.0%
ㅡ	5/5	5/5	100%
ㅣ	55/55	50/55	95.5%
ㅔ	20/20	20/20	100%
sp	20/20	19/20	97.5%
TOTAL	457/480 95.2%	449/480 93.5%	906/960 94.4%

(a)

음 소	화자 A	화자 B	인 식 율
ㄱ	16/20	18/20	85.0%
ㄲ	6/10	5/10	55.0%
ㄴ	5/10	10/10	75.0%
ㄷ	40/40	40/40	100%
ㅁ	61/65	52/65	86.9%
ㅅ	36/40	36/40	90.0%
ㅆ	8/10	10/10	90.0%
ㅇ	12/15	8/15	66.7%
ㅈ	5/10	6/10	55.0%
ㅋ	5/10	7/10	60.0%
ㅏ	107/110	107/110	97.3%
ㅓ	9/10	8/10	85.0%
ㅕ	5/5	4/5	90.0%
ㅗ	20/20	18/20	95.0%
ㅛ	4/5	5/5	90.0%
ㅡ	2/5	5/5	70.0%
ㅣ	54/55	55/55	99.1%
ㅔ	18/20	20/20	95.0%
sp	18/20	18/20	90.0%
TOTAL	431/480 89.8%	432/480 90.0%	863/960 89.9%

(b)

타를 확보하고 이 중 각 사람에게서 10개의 데이타 즉, 총 20개의 데이타를 신경망을 통해서 학습 시킨 후 학습에 참여한 20 개 문장 데이타 및 학습에 참여하지 않은 나머지 80 개 문장 데이타, 합해서 100 개 문장 데이타에 대해 음소 인식을 실시하고 인식 결과 나타난 음소열을 관찰열로 하여 $=\langle A, \pi\rangle$ 를 구하였다. 이 값으로 실험에 참여한 사람의 음성으로 10 개의 문장에 대해 각 10 번씩 각 사람에 대해 100 개 즉, 총 200개의 데이타를 새로이 확보하여 음소 인식을 실시하고 음소 인식된 결과를 후처리하여 각 문장에 대한 인식률을 조사하였다. 한편 초기상태 분포확률 값과 상태 천이확률이 0 이 나타나는 경우에는 이로 인해 문장 인식시에 각 모델의 확률이 0 이 되어 확률 계산의 의미가 상실되는 경우가 생기므로 이를 0 이 아닌 작은 값으로 수정해 주어야 한다. 본 시스템에서는 보정값이 0.001 일 때나 0.0001 일 때 가장 좋은 결과를 보이고 그 보다 더 크거나 작을 때 인식률이 저하됨을 확인하였다. 표 3은 문장 인식 결과로서 화자 A와 화자 B 모두 죄송합니다 에서 저조한 인식률을 나타내고 있다. 이것은 표 2의 (a)에서 보는 바와 같이 'ㅈ'의 인식률이 저조한 것과 밀접한 관련성을 나타내고 있다.

음소에 대한 인식률과는 달리 문장에 대한 인식의 경우 1 : 1 정합을 하면 한 음소라도 달리 나오면 인식을 못한 것으로 되기 때문에 인식률이 저조할 수 있다. 그러나 설사 신경망 출력에서 몇개의 음소를 제

표 3. 문장 인식 결과
Table 3. Sentence recognition results

음 성	화 자 A		화 자 B		인 식 율	
안녕하십니까	10/10		10/10		100%	
감사합니다	8/10		8/10		80.0%	
고맙습니다	8/10		10/10		90.0%	
계속합니다	9/10		10/10		95.0%	
그만합니다	10/10		10/10		100%	
어서오세요	10/10		10/10		100%	
믿겠습니다	10/10		10/10		100%	
시작합니다	9/10		10/10		95.0%	
사랑합니다	10/10		9/10		95.0%	
죄송합니다	7/10		7/10		70.0%	
TOTAL	91/100	91.0%	94/100	94.0%	185/200	92.5%

대로 인식하지 못할 경우가 생기더라도 후처리 과정의 Markov 모델에서 그런 경우가 모델링 되어 있기 때문에 인식에 성공하게 된다. 문장 인식률은 화자 A가 91% 화자 B가 94%를 얻어서 전체적으로 92.5%의 문장 인식률을 얻을 수 있었다.

Ⅵ. 결 론

본 논문을 통하여 연속 음성에서의 문장 인식 방법에 대하여 제안하였다. 25.6ms의 시간창을 사용하여 3ms마다 음성의 특징을 구하였으며, 음성의 시간 변화를 학습하기 위하여 넓은 구간의 시간 파형에 대하여 선택적으로 입력 벡터를 구성하였다. 사람의 음향처리 방법을 모방하여 사람의 귀의 특징을 고려한 7차 PLP 켑스트럼 계수를 특징값으로 사용하였다. 각 음성에 대하여 두 화자의 음성으로부터 학습 벡터를 추출하여 학습시켰다. 음소 인식 실험에는 총 100개의 음성을 사용하여 최대 94.4 %의 인식률을 얻었다.

문장에서 음소를 분리하여 학습 시킬 음소를 추출할 때 천이 구간을 제외시켜 신경망이 이 구간 데이타 때문에 학습에 혼선을 일으키지 않게 하였다.

문장 인식에 촛점을 맞추어 보면 인식된 음소열은 도중에 잘못 인식된 것이 포함되어 있으므로 Markov 모델을 이용하여 조금 잘못 인식된 음소열이 들어 있더라도 같은 문장으로 취급하여 이를 확률적으로 처리하였다. 이는 사람의 fault tolerance 기능을 모방한 것으로서 조금 잘못된 음소열도 훌륭히 인식하였다. 음소 인식 실험에 쓰였던 100개 문장을 Markov 모델을 구하는데 사용하였으며 여기에 참여하지 않은 200개의 음성 데이타로 문장 인식을 실시한 결과 92.5%의 문장 인식률을 얻을 수 있었다.

인공 신경망과 Markov 모델을 결합하여 문장 인식을 시도하였으며 94.4%의 음소 인식률을 보인 신경망으로부터 92.5%의 문장 인식률을 얻을 수 있었다. 후처리 과정으로서 Markov 모델 결합이 상당히 고무적인 결과를 가져 왔다. 신경망만 사용할 경우 다루기 힘든 후처리에서의 sequence 특성을 Markov 모델이 보완하였고 HMM을 사용할 경우 필요로하는 방대한 데이타의 부담을 신경망의 사용으로 해결하였다.

학습의 문제점으로서 문장 내에 존재하는 음소의 갯수가 음소마다 같지 않기 때문에 학습데이타의 불평등이 생기게 된다. 이런 문제를 해결한다면 음소 인식률의 향상과 함께 문장 인식률의 향상도 가져올 것으로 기대된다. 앞으로 대상 문장을 늘려서 실험해 볼 필요가 있으며 계속적인 연구 방향이 될 것이다.

참 고 문 헌

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, pp. 396-453, 1978, Prentice-Hall Inc.

2. S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*, 1985, Academic Press
3. T. W. Parsons, *Voice and Speech Processing*, pp. 59-81, 1986, McGraw Hill Inc.
4. P. D. Wasserman, *Neural Computing: Theory and Practice*, 1993, Van Nostrand Reinhold New York
5. Adam Blum, *Neural Networks in C++*, 1992, John Wiley & Sons, Inc.
6. Jacket M. Murada, *Introduction to Artificial Neural System*, pp. 163-250, 1992, WEST
7. L. R. Rabinar and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, 1993, Prentice Hall Inc.
8. H. Hamansky, "Perceptual Linear Predictive(PLP) analysis of speech" J. Acoust. Soc. Am., 87(4): 1738~1752, April 1990
9. Rik D. T. Janssen, Mark Fanty and Ronald, "Speaker Independent Phonetic Classification in Continuous English Letters," INNS vol. 2, pp. 801~808, 1991
10. H. Harmanskey, Kazuhiro Tsuga, Shozo Makino, and Wakita., "Perceptually Based Processing In Automatic Speech Recognition," ICASSP, pp. 1971-1162, 1986
11. L. R. Rabiner, S. E. Levinson and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition," The Bell System technical Journal, Vol. 62, No. 4, April 1983
12. L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 37, No. 8, Aug. 1989
13. M. R. Schroeder, "Recognition of Complex Acoustic Signals", Life Science Research Report 5, edited by T. H. Bullock (Abakon Verlag, Berlin), p. 324

▲김 선 일

1960년 3월 10일생

1983년 2월 : 아주대학교 전자공학과 (공학사)
1985년 2월 : 아주대학교 전자공학과 (공학석사)
1994년 2월 : 아주대학교 전자공학과 박사과정 수료
1985년 3월~1990년 2월 : 한국기계연구소 자동제어실 연구원
1990년 3월~1990년 8월 : 한국기계연구소 자동제어실 선임연구원
1990년 8월~현재 : 거제전문대 전자과 교수
※주관심 분야 : 음성인식, 인공지능, 신경회로망, 디지탈 신호처리

▲홍 기 원

1969년 1월 3일생

1994년 2월 : 아주대학교 전자공학과 (공학사)
1994년 3월~현재 : 아주대학교전자공학과 석사과정
※주관심 분야 : 디지탈 신호처리, 음성인식, 신경회로망

▲이 행 세

1943년 8월 29일생

1966년 2월 : 전북대학교 전기공학과 (공학사)
1972년 2월 : 서울대학교 전자공학과 (공학석사)
1984년 2월 : 고려대학교 전자공학과 (공학박사)
1968년~1970년 : 해군사관학교 전자공학 교관
1973년~현재 : 아주대학교 전자공학과 교수
1982년~1983년 : 미국 Columbia Univ. 객원교수
1987년~1988년 : 프랑스 INRIA 객원교수
1992년~1994년 : 거제전문대 학장
※주관심 분야 : 문자 및 음성인식, 인간-기계 인터페이스, 인공지능, 신경회로망, 디지탈 신호처리