論文95-32B-12-10

제한된 한국어 연속음성에 나타난 음소인식에 관한 연구

(A Study on the Phoneme Recognition in the Restricted Continuously Spoken Korean)

沈 成 龍 *, 金 善 一 **, 李 幸 世 **

(Sungryong Shim, Seonil Kim, and Haing Sei Lee)

요 약

본 논문은 한국어 연속음성의 기계인식 알고리듬에 관한 것이다. 제안된 알고리듬은 고정구조 인공신경망을 사용한다. 인공신경망의 입력으로는 영교차율, 단구간 에너지 그리고 청각특성을 고려한 인지선형예측계수(PLP) 혹은 비교를 위하여 PARCOR 계수 등을 사용하며 171ms의 시간구간을 포함한다. 인지선형예측계수(PLP)를 사용한 프레임 기반의 음소인식실험은 소규모 단어인식 실험에 대하여 약 99%의 인식률을 보였다. 결과적으로 제안된 PLP 방법이 음소의 인식에 적합함을 보였다.

Abstract

This paper proposes an algorithm for machine recognition of phonemes in continuously spoken Korean. The proposed algorithm is a static strategy neural network. The algorithm uses, at the stage of training neurons, features such as the rate of zero crossing, short-term energy, and either PARCOR or auditory-like perceptual linear prediction(PLP) but not both, covering a time of 171ms long. Numerical results show that the algorithm with PLP achieves approximately the frame-based phoneme recognition rate of 99% for small vocabulary recognition experiments. Based on this it is concluded that the proposed algorithm with PLP analysis is effective in phoneme recognition.

I. 서 론

사람이 사용하는 언어에 있어서 단어의 수 만도 수십만 가지이며 지역이나 사람에 따라 또는 어조에 따라 음향학적인 음성의 특징은 천차만별이다. 이러한 음성은 기계가 인식하기에 매우 복잡하고 어려운 일이며, 이 때문에 제한적으로 인식을 하고자 하는 시도가 많이 있어 왔다[12]. 음절단위의 인식은 인식해야할 대상의 수를 매우 줄여주기는 하지만 훈련된 사람만이 인식기와 대화할 수 있는 단점을 안고 있다[12]. 또한 단어 단위의 인식은 그 응용이 극히 제한적이며 응용의 폭을 넓히려고 하는 경우, 대규모의 데이터베이스를 필요로 하며 대단히 많은 연산량을 가진다[12]. 언어를 음소단위로 인식하는 것은 가장 자연스럽고 데이터베이스를 가장 적게 운용할 수는 있으나 음소간의 연음현상이 음소를 음향학적으로 전혀 다른 것으로 만들므로 모든 경우를 인식하기가 어려우며, 더우기 사람의 어조에 따른 변화는 해결하기 대단히 어려운 문제가 된다.

본 논문에서는 이러한 문제를 해결하기 위하여 두 가지 측면에서 접근을 시도하였다. 비슷한 음성에 대하여 비슷한 특성을 가지는 파라미터의 추출을 위하여 사람의 청각신경을 모방한 인지선형예측법(Percep-

* 正會員, 高等技術研究員
(IAE)
** 正會員, 亞洲大學校 電子工學科
(Dept. of Elec. Eng. , Ajou Univ.)
接受日字:1995年1月3日, 수정완료일:1995年11月20日

tual Linear Prediction)을 적용하여 화자 독립적인 음성의 특징을 구하고자 하였고, 분석구간에 인접한 구간에 대해서도 입력으로 사용함으로써[14] 인접한 음소에 의한 음소의 변화를 학습함으로써 화자마다 다른 발음습관에 의한 변화를 흡수하고자 하였다.

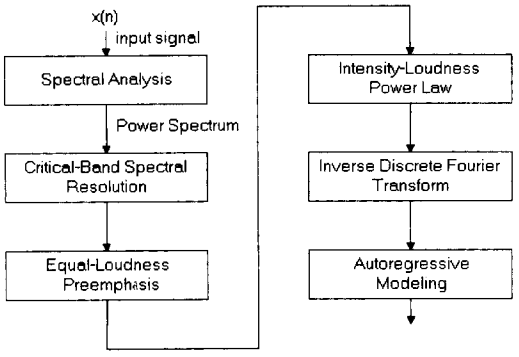

그림 1. PLP 분석법의 블럭도
Fig. 1. Block diagram of PLP speech analysis method.

Ⅱ. 인지선형예측 분석법

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[8]. 또한 음성 스펙트럼의 전력 변화율을 감소시키기 위하여 3 제곱근 처리를 함으로써 intensity-loudness 3 제곱근 크기 압축을 실시한다[8]. 이러한 처리를 거친 16 개의 스펙트럼 성분에 퓨리에 역변환 과정을 적용시켜 자기 상관 계수[6,7]를 얻는다. 전극점 모델은 얻어진 자기상관 계수로부터 원하는 차수로 계산되며, 이로부터 다시 캡스트럼 계수를 계산할 수 있다. 본 논문에서는 7 차의 전극점 모델을 사용하였다.

1. 주파수 분석(Spectral Analysis)

음성신호의 세그먼트는 해밍창(Hamming Window)에 의해 처리된다.

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

단, N은 창의 길이이다. 전형적인 창의 길이는 20ms이나 FFT(Fast Fourier Transform)를 위하여 25.6ms(256 point)의 시간창을 사용하였으며, 단구간 제곱합을 구하기 위하여 전력 스펙트럼(power spectrum)은 식(2)와 같이 실수성분과 허수성분을 제곱하여 더한다.

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

2. 임계대역 주파수 해법(Critical-Band Spectral Resolution)

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)에 의해 제안되었다. 결과치인 굴절된 전력스펙트럼(power spectrum)은 임계대역 마스킹 곡선(critical-band masking curve) $\Psi(\Omega)$와 콘볼루션(Convolution)된다. 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(discrete) 콘볼루션(convolution)은 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(critical-band masking curve) $\Psi(\Omega)$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 여파기들을 나타낸다. 16개의 여파기들이 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.

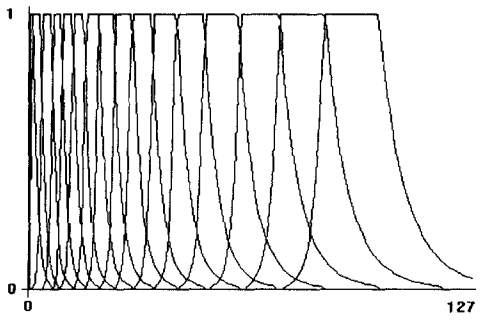

그림 2. 임계대역주파수해법에 의한 여파기
Fig. 2. The filters from critical-band spectral resolution.

3. Equal-loudness pre-emphsis

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

$$\Xi[\ \Omega(\omega)] = E(\omega)\Theta[\ \Omega(\omega)] \tag{6}$$

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

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

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

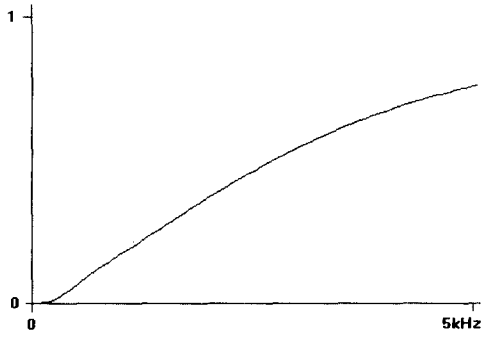

그림 3. Equal-loudness pre-emphsis 여과기의 특성
Fig. 3. The characteristics of equal-loudness pre-emphasis filter.

식 (7)은 0과 400Hz 사이에서 12dB/oct, 400과 1200Hz 사이에서 9dB/oct, 1200과 3100Hz 사이에서 6dB/oct 그리고 3100Hz 이상에서 0dB/oct의 감쇄율을 가지는 전달함수를 표현한다. 보통의 소리 수준에서는 이러한 근사방법이 5000Hz까지 매우 정확하다. 하지만 더 높은 대역폭을 갖는 응용에 있어서는 5000Hz 이상에서 더 급격한 감쇄율을 갖는 항을 첨가하는 것이 유용하다. 이것을 다음에 나타내었다.

$$E(\omega) = \frac{(\omega^2+56.8\times10^6)\omega^4}{(\omega^2+6.3\times10^6)^2(\omega^2+0.38\times10^9)(\omega^6+9.58\times10^{26})} \tag{8}$$

4. Intensity-loudness power law

전극점 모형화(all-pole modeling) 전의 처리단계 중의 마지막은 3 제곱근 크기 압축(cubic-root amplitude compression)이다.

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \tag{9}$$

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

5. 자기회귀 모델링(Autoregressive modeling)

처리과정을 거친 스펙트럼은 결과적으로 전력 스펙트럼이므로 퓨리에 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 사용하였다. 이와 같이 얻어진 자기상관 계수들은 전극점 모델을 구하는 자기회귀처리에 직접 이용할 수 있으며 또한 켑스트럼 계수와 같이 다른 계수를 구하는데 이용할 수 있다.

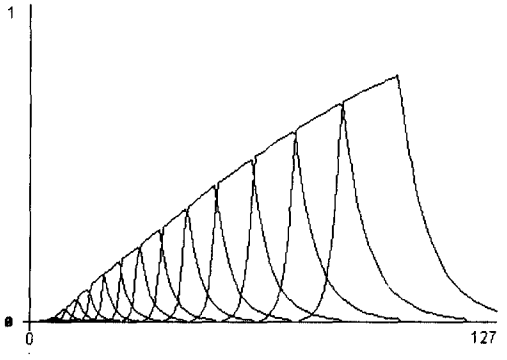

그림 4. 16개의 가중함수
Fig. 4. The 16 weighted functions.

6. 실제적인 고려

실제적으로, 콘볼루션(convolution)과 프리엠퍼시스(Pre-emphasis)는 $\Xi(\Omega)$의 각각의 표본에 대하여 $\Xi(\Omega)$ 표본 하나당 각각의 가중된 $P(\omega)$ 표본의 스펙트럼의 합으로 표현될 수 있다. 즉,

$$\Xi[\ \Omega(\omega_i)] = \sum_{\omega=\omega_n}^{\omega_n} w_i(\omega)P(\omega) \qquad (10)$$

식(10)은 식(4), (5)을 이용하여 얻어지며, 식(3)을 역으로 풀어서 다음을 얻는다.

$$\omega = 1200\pi\sinh(\frac{\Omega}{6}) \qquad (11)$$

그림 4에 $w_i(\omega)$의 특성에 대하여 도시하였다.

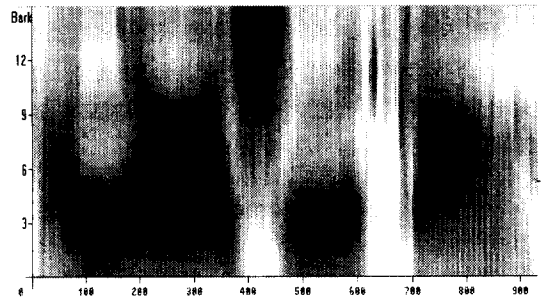

그림 5. PLP계수에 의한 스펙트로그램("안녕하십니까")
Fig. 5. The spectrogram from PLP coefficients.

PLP분석을 위한 연산 복잡도는 LP분석에 비해 대단히 크다. 연산측면에서 가장 복잡한 부분이 FFT 스펙트럼 계산과, 이어지는 임계대역 스펙트럼 적분과 3제곱근 압축등이다. AR 모델을 위한 연산의 복잡도는 낮아진 주파수 분해능으로 인하여 무시할 수 있을 정도로 작다. 그림 5 에 인삿말 "안녕하십니까"의 PLP에 의한 스펙트로그램(spectrogram)을 도시하였다. 가로축은 시간을 나타내며 세로축은 Bark-주파수를 나타낸다. 더 짙게 표시된 부분의 주파수 성분이 더 큰 에너지를 가진다.

PLP 모델의 결과 스펙트럼은 일반 LP 모델의 스펙트럼에 비해 더 선형적이다[15]. 또한 일반 LP 모델에 비해 더 낮은 차수의 모델링이 가능하다[15]. 결과적으로 인공신경망의 입력의 감소와 데이타베이스의 역할을 하는 가중치들을 줄이는 역할을 하게되며, 이것은 처리속도의 효율화에 기여하게 된다[10]. 또한 선형적인 특징들은 화자가 발음한 음성특징중 화자 종속적인 부분을 상당히 감소시킨다고 생각되며 실험을 통하여 이를 확인하였다.

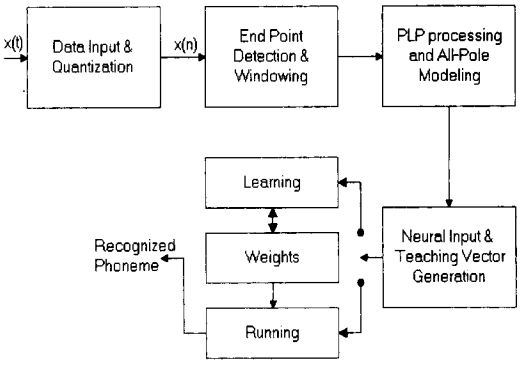

그림 6. 제안된 시스템의 전체 블럭도
Fig. 6. The block diagram of proposed system.

Ⅲ. 제안하는 인식시스템

1. 전체 시스템의 구성

시스템은 전체 3개의 부분으로 구성되어 있다. 데이타의 입력과 편집을 위한 부분과, 데이타의 처리를 위한 부분 그리고 인식기의 학습과 인식을 위한 부분이 그것이다. 각각에 대하여 다음에 설명하였다.

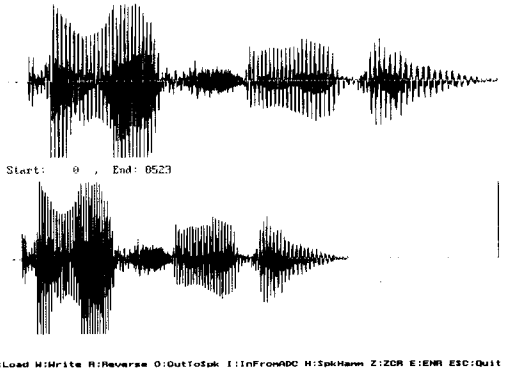

그림 7. 파형의 편집화면
Fig. 7. The screen of waveform editing.

2. 데이터의 입력과 편집[3,4,5]

PC486하에서 DT2801A 보드를 사용하여 10kHz 12bit 양자화 하였다. 데이터의 입력을 위하여 프로그램은 단구간 에너지를 이용하여 음성의 시작점을 자동으로 검출하도록 작성되었고 음성의 시작점으로부터

1.6초간 입력을 받을 수 있다. 또한 입력된 음성을 적당한 크기로 잘라 텍스트 파일로 저장할 수 있다. 입력과 편집화면을 그림 7에 도시하였다.

3. 데이타의 처리

1) 인지선형예측 처리 및 전극점 모델링[13,14]

음성 표본을 3ms마다 256 표본을 취하여 PLP처리를 하고 전극점 모델을 얻는다. 얻어진 계수들은 영교차율, 단구간 에너지등과 함께 인식기의 입력을 구성한다.

▶ **PLP 켑스트럼 계수**[7] : 제 7 차 인지 선형 예측 계수를 이용한 켑스트럼 계수이다. PLP를 이용한 켑스트럼 계수는 PLP 계수로부터 다음과 같이 얻을 수 있다.

$$C_i = a_i + \sum_{j=1}^{i-1} C_j \cdot a_{i-j} \qquad i=1,2,\cdots,p \quad (12)$$

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

$$\hat{C}_i = C_i \cdot (i+1) \qquad i=1,2,\cdots,p \quad (13)$$

▶ **단구간 에너지**[7] : 25.6ms의 음성 신호에 해밍 윈도우 처리를 한 제곱 합이며, PLP 처리과정에서 상당부분 왜곡된다.

▶ **단구간 영교차율**[7] : 단위 시간에 0을 지나는 횟수를 말하며 256 표본의 음성신호중 0을 지나는 횟수를 계수 하여 사용하였다.

2) 학습을 위한 학습패턴의 구성

학습패턴은 인식기의 학습을 위한 목표 값이며 올바른 목표치를 사용하여야 올바르게 학습을 시킬 수 있다[9,10]. 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[9,10]. 그러므로 다음과 같은 방법으로 올바른 목표치를 정하고 음소의 경계에 대하여 일정구간 학습을 금지시킴으로써 목적을 달성할 수 있다. 이와 같은 구성은 음소경계에 대한 애매한 출력, 즉 경계 양쪽 음소의 특징을 모두 포함하는 출력을 내도록 인식기를 유도할 것이며, 이것이 바람직한 출력이다.

그림 8 은 목표치를 설정하는 방법을 나타낸 것이다. 음영으로 나타낸 부분은 주위 음소의 학습에 영향을 미치지만 자신은 학습되지 않는다. 각각의 음소의 길이는 단어에 따라 달라지며 음영으로 표시된 부분은 20ms로 임의로 설정하여 학습을 하였다.

그림 8. 학습을 위한 목표치의 설정
Fig. 8. The setting of teaching vector for learning.

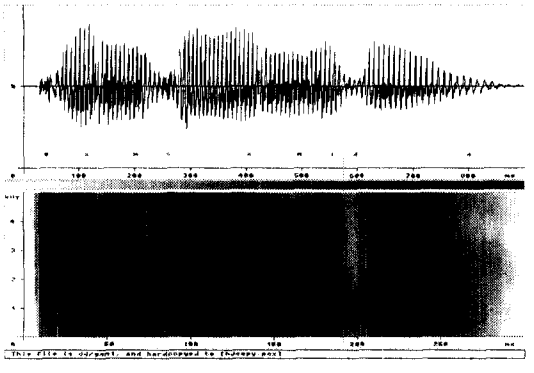

그림 9. 목표값 생성을 위한 입력방식
Fig. 9. The input method for teaching vector generation.

학습을 위한 목표치의 입력은 파형과 스펙트럼이 동시에 표시된 화면상에서 행해진다. 화면을 주시하면서 키보드를 이용하여 세로막대 커저(cursor)를 이동시킨 후 순차적으로 목표 값을 입력하면 프로그램은 좌표와 목표 값을 입력받아 해당하는 위치의 프레임들에 대하여 입력 값과 목표치를 파일에 기록한다. 이것을 그림 9에 나타내었다.

4. 학습 및 인식

인식기로서 사용되어지는 인공신경망은 오차역전달 알고리듬을 이용한 다층구조(MLP) 신경망을 사용하였으며, 인공신경망의 학습 및 인식을 위하여 인공신경망의 입력단 및 학습 벡터의 구성을 다음과 같이 하였다.

1) 입력벡터의 구성[14]

연속음에 나타나는 음소는 전후의 다른 음소와의 조음 현상에 의한 변화가 심하다. 현재의 음성 조각을 인식하기 위해서는 인접한 음소를 같이 참조하는 것이 타당하며, 더더욱 사람의 그것과 유사해진다. 보통은 이러한 변화를 학습하기 위하여 시간지연신경망을 사용하지만 본 논문에서는 고정 신경망을 사용하고, 입력

단의 구성을 달리하여 시간 변화를 학습하였다.

3ms의 음성 프레임을 인식하기 위하여 특징들의 집합을 구성한다. 집합은 총 63개의 입력으로 구성되며, 각각의 요소에 대하여 그림 10에 도시하였다.

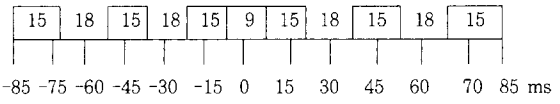

그림 10. 입력벡터의 요소 구성
Fig. 10. The distribution of 63 PLP features.

음성은 3ms 마다 25.6ms 에 대하여 특징이 추출되며 전체 171ms 의 특징들에 대하여 그림 10에 의해 선택적으로 추출된 특징들을 평균하여 1블럭당 9개씩, 7개의 블럭에 대해 63개의 특징 파라메타가 구해진다.

실선은 PLP가 평균되어지는 구간의 길이를 나타내며, 점선은 사용되지 않는 구간의 길이이다. 현재의 프레임에 대하여 인접한 3개의 프레임(9ms)을 평균하였으며, 나머지는 인접한 다섯 개의 프레임(15ms)에 대하여 평균을 취하였다. 결과적으로 전체 7개의 블럭으로 구성되며 각각의 블록은 9개의 데이터로 구성되어 있는데, 7차의 PLP 계수와 1개의 단구간 에너지 그리고 1개의 영교차율이 그것이다.

2) 학습 및 인식

음소를 인식하는 알고리듬은 음성의 입력을 한국어의 음소를 가리키는 32개의 인공신경망의 출력으로 내보낸다. 이러한 음소들은 "ㄱ", "ㄲ", "ㄴ", "ㄷ", "ㄸ", "ㄹ", "ㅁ", "ㅂ", "ㅃ", "ㅅ", "ㅆ", "ㅇ", "ㅈ", "ㅉ", "ㅊ", "ㅋ", "ㅌ", "ㅍ", "ㅎ"의 19개의 자음과, "ㅏ", "ㅑ", "ㅓ", "ㅕ", "ㅗ", "ㅛ", "ㅜ", "ㅠ", "ㅡ", "ㅣ", "ㅚ", "ㅐ" 등의 12개의 모음과 묵음을 나타내는 "space" 1개의 총 32개로 구성되었다. 매 3ms마다 25.6ms의 시간영역 신호에 대하여 특징을 추출하고, 학습 및 인식을 실시하였다.

Ⅳ. 실험 및 검토

1. 실험 환경

한국어 음소의 인식을 위하여 PLP 계수를 주요 특징으로 사용하였고, 이의 타당성을 검증하기 위하여 PARCOR 계수를 사용한 시스템과 비교하였다. 대상 음성으로는 2개의 한국어 인사말("안녕하십니까", "감사합니다")을 사용하였고, 한사람이 발음한 음성만으로 학습한 후 다른 4사람의 음성에 대하여 인식 실험을 행하였다.

데이터의 처리를 위하여 PC용 UNIX인 Linux 환경하에서 GNU-C를 사용하여 프로그램이 작성되었다. UNIX 기반 OS인 Linux를 사용하는 잇점은, 일반적으로 신호처리가 대규모의 메모리를 요구하므로, 거의 무제한의 메모리(Virtual Memory)를 사용할 수 있다는 점과 고해상도의 그래픽을 사용할 수 있다는 데에 있다. 또한 빠른 처리속도와 MS-DOS와의 데이터 호환성도 무시할 수 없는 잇점이다. 그래픽의 처리를 위하여 함께 제공되는 VGA 라이브러리를 사용하였고, 논문 전반에 걸쳐 여기에서 처리한 결과를 도시하였다. 마지막 단계인 인식 및 학습기는 처리속도의 향상을 목적으로 워크스테이션 상에서 작성하였다.

2. 인식 여부의 판단

연속음중에 나타나는 음소들은 인접 음소와의 연음현상으로 인하여 구분하기 어려운 경우가 대부분이며 음소들이 겹쳐진 경우가 많아서 인식 결정이 대단히 어렵다. 따라서 판단을 위하여 다음과 같은 몇가지 기준이 마련되었다.

- ▶ 두가지의 출력이 동시에 활성화 된 경우 활성화 정도가 큰 출력이 승자가 된다.
- ▶ 잠깐 동안만 활성화 되는 출력에 대해서는 인접한 출력들의 영향으로 제거된다.
- ▶ 부분적인 오인식은 중간정도의 인식 성공으로 인정한다(하나의 음소구간 동안 절반 이하는 인식은 인식되지 않고, 나머지의 경우에는 판단 알고리듬의 구현방법에 따라 상황이 달라질 수 있으므로 중간정도의 인식 성공으로 인정한다)

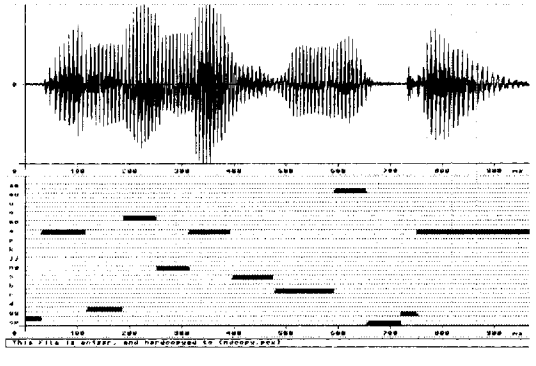

그림 11. 인공신경망의 인식결과
Fig. 11. The recognition results of neural network.

그림 11은 인식실험에 참여한 음성의 파형과 인식기의 출력을 같은 시간축상에 나타낸것이다. 그림에 나타낸 것 처럼 세로축은 시간, 즉 분석 구간을 나타내며 가로축은 인식기의 출력을 순서대로 나타낸 것이다. 각각의 세로칸은 해당하는 음소와의 연관 정도에 대한 인식기의 출력 강도이다. 가장 밑에 나타난 것은 space로서 묵음 구간을 나타낸다. 그 다음이 'ㄱ'과, 그 다음의 'ㄲ'등의 순서로 가장 마지막이 'ㅐ ㅖ'이다. 사람이 분간하기 어려운 음소는 음향학적으로도 별 차이가 없으므로 'ㅐ'와 'ㅖ' 등은 같은 것으로 처리하였다. "안녕하십니까"에 나타나는 'ㅎ'음은 시간파형이나 스펙트럼상에서는 'ㅇ'와 동일하게 나타나는 관계로 본 논문에서는 'ㅎ'음의 인식을 실시하지 않았다. 또한 뒤에 나타나는 'ㄴ'은 'ㅁ'의 영향으로 발음하는 사람의 습관에 따라 "안녕하심미까"로 되는 경우가 많고 나타나더라도 구분하기가 어려운 경우가 대부분이었다. 따라서 이 역시 본 논문의 실험에서 제외되었다.

3. 실험결과 및 검토

제안된 방법의 타당성을 입증하기 위하여 비교적 성능이 우수하다고 알려진 PARCOR 계수를 사용한 시스템과 동일 조건에서 비교하였다. PLP와 PARCOR의 차수는 공히 7차의 계수를 사용하였고, 영교차율과 단구간 에너지를 동시에 사용하였다. 또한 PARCOR 계수에 있어서 최적이라고 알려진 12차에 대하여도 동일한 실험을 하여 비교하였다. 시간 변화를 관찰하는 등의 인공신경망의 데이타 입력방법 또한 동일하다. 실험은 두가지에 대하여 행하였으며 첫번째는 단일 화자가 발음한 5개씩의 음성을 학습에 사용하여 비교하였다. 두번째는 5명의 화자가 각각 1번씩 발음한 음성을 학습에 사용하여 학습에 참여하지 않은 음성을 사용하여 실험하였다.

학습과 인식을 위한 특징의 추출은 3ms 마다 추출되므로 900ms 의 인삿말에 대해 대략 300 개 정도의 벡터가 생성된다. 본 실험에서는 사용한 데이터는 5개씩 2 종류를 사용하였으므로 10 개의 단어가 되며 이는 3000개 정도의 벡터를 생성시킨다. 프로그램은 학습 및 인식시에 인접한 57 개의 벡터에 대하여 부분적인 평균을 계산하도록 작성되었으며, 메모리의 사용효율을 높이는 대신에 연산시간이 다소 길어졌다.

■ 각 음소별 인식 결과

각 단어별 또는 화자별 인식률은 지면 관계상 생략하였으며, 다음에 각 인식시스템에서의 음소별 인식률을 나타내었다.

▶ 단일화자 학습시스템의 인식 결과

한사람이 발음한 다섯개씩의 음성을 대상으로 학습한 시스템에서 학습에 참여한 1인과 참여하지 않은 4인이 발음한 음성에 대한 인식률을 음소별로 나타내었다. PLP를 이용한 시스템과 PARCOR를 이용한 시스템을 동시에 나타내었다. 사람에 따라 각각 다른 인식률을 나타내었으며 특정 화자에 대하여 특히 나쁜 인식률을 나타내었다. 이는 인식기가 화자마다 다른 발음습관을 학습하지 못한 때문으로 해석되어진다. 표 1에 나타낸 바와 같이 PLP를 이용한 시스템이 PARCOR를 이용한 시스템에 비해 좋은 결과를 보였으며 PLP가 화자 독립적인 음소의 특징을 더 잘 표현한다는 것을 나타낸다.

표 1. 단일화자 학습에 대한 음소별 인식결과

Table 1. The experimental results of recognition rate of phonemes for single speaker teaching.

분석법 / 음소	PLP (단위:개)		PARCOR (단위:개)	
	인식갯수	전체갯수	인식갯수	전체갯수
ㄱ	22.5	25	20.5	25
ㄲ	16.0	25	7.5	25
ㄴ	19.0	25	11.5	25
ㄷ	25.0	25	10.5	25
ㅁ	66.0	75	66.0	75
ㅅ	44.5	50	45.0	50
ㅇ	17.5	25	15.0	25
ㅏ	150.0	150	147.5	150
ㅖ	19.5	25	19.0	25
ㅣ	48.5	50	41.5	50
sp	25	25	20.0	25
평 균	90.7(%)		80.8(%)	

▶ 복수화자 학습시스템의 인식결과 1

5명이 발음한 음성중 각각 1개 음성을 학습대상으로 하였으며 나머지에 대하여 인식률을 산출하였다.

표 2에 나타낸 것과 같이 PLP가 더 우수한 성능을 나타내었다. 여러사람으로부터의 음성을 학습하였기 때문에 음소에 대한 더 일반적인 특징들이 학습되었으며, 높은 인식률을 나타내었다.

표 2. 복수화자 학습에 대한 음소별 인식결과 1

Table 2. The experimental results 1 of recognition rate of phonemes for multiple speaker teaching.

분석법 / 음소	PLP (단위:개)		PARCOR (단위:개)	
	인식갯수	전체갯수	인식갯수	전체갯수
ㄱ	25.0	25	23.0	25
ㄲ	24.5	25	22.5	25
ㄴ	20.5	25	19.0	25
ㄷ	25.0	25	14.5	25
ㅁ	74.0	75	70.0	75
ㅅ	50.0	50	47.5	50
ㅇ	25.0	25	18.0	25
ㅏ	150.0	150	146.0	150
ㅕ	24.5	25	19.5	25
ㅣ	50.0	50	47.0	50
sp	25.0	25	25.0	25
평 균	98.7(%)		90.4(%)	

▶ 복수화자 학습시스템의 인식결과 2

PLP에 대하여는 인식결과 1과 동일하며 PARCOR에 대하여 12차를 적용하여 비교하였다.

표 3. 복수화자 학습에 대한 음소별 인식결과 2

Table 3. The experimental results 2 of recognition rate of phonemes for multiple speaker teaching.

분석법 / 음소	PLP (단위:개)		PARCOR (단위:개)	
	인식갯수	전체갯수	인식갯수	전체갯수
ㄱ	25.0	25	24.0	25
ㄲ	24.5	25	9.0	25
ㄴ	20.5	25	15.0	25
ㄷ	25.0	25	14.0	25
ㅁ	74.0	75	74.0	75
ㅅ	50.0	50	49.5	50
ㅇ	25.0	25	20.0	25
ㅏ	150.0	150	148.5	150
ㅕ	24.5	25	21.5	25
ㅣ	50.0	50	45.5	50
sp	25.0	25	24.0	25
평 균	98.7(%)		89.0(%)	

12차의 PARCOR계수를 사용한 시스템의 인식률이 7차의 PLP를 사용한 시스템의 인식률 보다 더 낮게 나타난 것은 물론이고 7차의 PARCOR를 사용한 시스템의 인식률보다 더 낮게 나타났다. 특히 비음(ㄱ,ㅁ,ㅇ)에 대한 인식률이 상대적으로 떨어지는 결과를 보였다. 높은 차수의 계수가 음성을 보다 더 정확하게 표현하며 이것 때문에 인공신경망이 화자 독립적인 음성의 특징을 잘 학습하지 못했다고 생각된다.

V. 결 론

본 논문을 통하여 연속음성에서의 음소를 인식하는 방법에 대하여 제안하였다. 25.6ms의 시간창을 사용하여 3ms마다 음성의 특징을 구하였으며, 음성의 시간변화를 학습하기 위하여 171ms의 시간파형에 대하여 선택적으로 입력벡터를 구성하였다. 음성의 특징을 표현하기 위하여 사람의 귀의 특징을 고려한 PLP와 일반적으로 성능이 우수하다고 알려진 PARCOR계수를 각각 사용하여 비교하였다. 각 음성에 대하여 한 화자의 음성으로부터 학습벡터를 추출하여 학습하였으며 학습에 참여한 사람을 포함하여 5명의 화자가 발음한 2가지의 연속음에 대하여 실험하였다. 단일화자가 발음한 음성으로 학습한 시스템에서 PLP를 이용하여 평균 90.7%의 인식률을 얻었으며, PARCOR를 이용하여 평균 80.8%의 인식률을 얻었다. 복수화자가 참여한 시스템에서 PLP를 이용한 시스템이 98.7%로 PARCOR를 이용한 시스템 보다 높은 인식률을 얻었다. 결과적으로 제안된 방법이 음성을 음소단위로 인식하는데 효과적이고, PLP 계수가 화자독립적인 음소의 특징을 잘 표현한다는 사실을 실험을 통하여 확인하였다.

앞으로 연구할 과제로는 음소에 대한 연구가 선행되어야 하며, 그후 더 많은 데이타로 부터 일반적인 음소의 특징을 학습하여 가능한 모든 음소에 대하여 적절한 출력을 낼 수 있도록 유도하는 것이다. 또한 출력단에 단어사전등을 구성하여 출력을 보정함으로써 단어에 대해 더 높은 인식률을 보장하고 훼손된 정보에 대한 복구를 시도할 수도 있을 것이다.

참 고 문 헌

[1] 이재건 외, "음소인식에 의한 한국어 단음절 인

식", 1992. 대한전자공학회 추계종합학술대회 논문집 제15권 제2호 pp. 665~669

[2] 이욱재 외, "우리말 복모음인식에 관한 연구", 1993, 제3호 인공지능, 신경망 및 퍼지시스템 종합학술대회논문집 pp. 217~219

[3] 심성룡 외, "BP알고리듬을 사용한 한국지명음성의 인식방법", 1993, 대한전자공학회 추계종합학술대회논문집 제16권 제2호 pp.992~995

[4] 이재건, 인공신경망을 이용한 한국어 단어음성 인식에 대한 연구, 1994, 석사학위논문 아주대학교

[5] 심성룡 외, "인공지능 방법을 이용한 음성의 피치검출 방법에 관한 연구", 1994, 대한전자공학회 추계종합학술대회논문집 제17권 제1호 pp. 845~847

[6] L. R. Rabinar and R. W. Schafer, *Digtal Processing of Speech Signals*, pp. 396-453, 1978, Prentice-Hall Inc.

[7] S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*, 1985, Academic Press.

[8] T. W. Parsons, *Voice and Speech Processing*, pp. 59-81, 1986, McGraw Hill Inc.

[9] P. D. Wasserman, *Neural Computing: Theory and Practice*, 1993, Van Nostrand Reinhold New York.

[10] Adam Blum, *Neural Networks in C++*, 1992, John Wiley & Sons, Inc.

[11] Jacket M. Murada, *Introduction to Artificial Neural System*, pp. 163-250, 1992, WEST.

[12] L. R. Rabinar and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, 1993, Prentice Hall Inc.

[13] H. Hamansky, "Perceptual Linear Predictive(PLP) analysis of speech" J. Acoust. Soc. Am., 87(4):1738~1752, April 1990.

[14] Rik D. T. Janssen, Mark Fanty and Ronald, "Speaker Independent Phonetic Classification in Continuous English Letters", INNS vol. 2, pp.801 ~808, 1991.

[15] H. Harmanskey, Kazuhiro Tsuga, Shozo Makino, and Wakita., "Perceptually Based Processing In Automatic Speech Recognition", ICASSP, pp. 1971-1162, 1986.

저 자 소 개

沈 成 龍(正會員)

1990년 2월 아주대학교 전자공학과(학사). 1995년 2월 아주대학교 전자공학과(석사). 1995년 3월 ~ 현재 대우고등기술원 근무

金 善 一(正會員)

1983년 2월 아주대학교 전자공학과(학사). 1985년 2월 아주대학교 전자공학과(석사). 1994년 2월 아주대학교 전자공학과 박사과정 수료. 1985년 3월 ~ 1990년 8월 한국기계연구소 근무. 1990년 8월 ~ 현재 거제전문대 전자과 조교수

李 幸 世(正會員) 第 32卷 B編 第 1 號 參照

현재 아주대학교 전자공학과 교수