'94 대한전기학회 추계학술대회 논문집(1994. 11. 18)

Cepstrum 계수와 Frequecy Sensitive Competitive Learning 신경회로망을 이용한 한국어 인식.

°이 수 혁 , 조 성 원 , 최 경 삼
홍익대학교 전기·제어공학과

Korean Digit Recognition Using Cepstrum coefficients and Frequency Sensitive Competitive Learning

°Suhyuk Lee , Seongwon Cho , Gyung-sam Choi
Dept. of Electrical & Control Engineering,
Hong-Ik University

ABSTRACT

In this paper, we pesent a speaker-dependent Korean isolated digit recognition system. At the preprocessing step, LPC cepstral coefficients are extracted from speech signal, and are used as the input of a Frequency Sensitive Competitive Learning(FSCL) neural network. We carried out the postprocessing based on the winning-neuron histogram. Experimetal results indicate the possibility of commercial auto-dial telephones.

1. 서론

근래에 와서 컴퓨터 하드웨어의 발전과 그로인한 성능의 향상으로 인해 최근에는 좀더 인간에게 친숙한 Interface 방식이 연구되어 왔다. 음성인식과 문자인식 등이 대표적인 방법으로, 그 중 음성인식에 관한 연구는 70년대 부터 활발히 진행되었으나, 타 분야에 비해 진척이 늦은 것은 음성 자체의 다양성에 기인한다. 즉 화자의 심리적, 신체적 상태, 주위 환경, 음성패턴의 불확실성, 음성의 중첩이나 왜곡 등이라 할 수 있다. 이러한 문제를 해결하기 위해 80년대 중반부터 Vector Quantization, Hidden Markov Model 등의 연구가 주종을 이루어 왔었다. 최근에는 신경회로망 이론과 퍼지이론의 연구가 활발히 진행되고 있으며, 각각 인간의 인지능력과 판단 방법을 모방한 것임을 볼 때 두 이론이 음성인식에 적용되는 것은 당연하다 하겠다 [8][9][10].

본 논문은 Frequency Sensitive Competitive Learning(FSCL)을 이용한 한국어 숫자음인식에 관한 연구이며, Simulation을 통해 자동다이얼 전화기 개발의 타당성 검증을 목적으로 한다. 다루는 범위는 화자종속, 고립단어, 한정단어 인식이다. 이때 인식단어는 한국어 숫자음 10개이고, 전처리는 인식률이 비교적 높은 Linear Predictive Coding(LPC)을 통한 Cepstrum 계수법을 사용하였고 [7], FCSL의 승자뉴런 히스토그램에 의한 후처리 부를 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 음성의 전처리에 대하여 설명하고, 3장에서는 FSCL신경회로망 이론을 다룬다. 4장에서는 후처리부, 그리고 5장에서는 실험결과를 살펴본다. 끝으로 6장은 결론이다.

2. 음성신호의 특징 추출

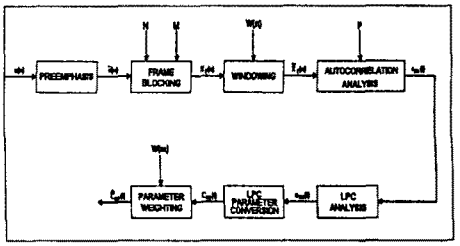
음성신호의 특징 추출방법으로 자주 사용되는 LPC법은 인간의 발성기관을 모델링하고, 그를 통해 특징 파라메터를 구하는 방법이며, 현재의 음성샘플은 과거 유한개의 샘플의 선형결합으로 표현할 수 있다는 것이 그 개념이다[4]. 이러한 개념에 의한 성도의 모델은 (식 1)과 같이 All Pole Model로 표현된다.

(식 1)의 LPC 계수로 부터 Cepstral 계수를 구하고 적절한 Weighting을 가하여 특징 벡터로 사용하면 잡음에 강한 음성특징을 구할 수 있다. cepstral 계수는 주파수 영역 신호의 위상 성분을 제거하고, 퓨리에 역변환을 통해 구해진다. 이 때 변환된 영역을 Quefrency 영역이라 한다 [1]. 잡음 부분은 고역 quefrency 영역으로 분리되어 쉽게 제거될 수 있다.

$$H(z)=\frac{S(z)}{U(z)}=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}} \qquad \text{(식 1)}$$

$H(z)$: 성도의전달함수
$S(z)$: 음성출력
$U(z)$: 성도의입력
G : *gain*
a_k : *filter*계수
p : *LPC* 차수

(그림 1)은 음성신호의 LPC 해석을 통한 cepstrum 계수를 구하는 블럭도이다 [5].



(그림 1) LPC 해석을 통한 Cepstrum 처리 블럭도

음성신호의 특성상 고주파 대역의 에너지가 저주파에 비해 작으므로 평탄한 spectrum을 갖게하기 위해 (식 2)의 필터를 통과시켜 Preemphasis한다.

$$H(z)=1-\tilde{a}z^{-1}, \qquad 0.9\le a\le 1.0 \qquad \text{(식 2)}$$

그 다음 과정으로 음성 주파수가 안정된 약 30mS 정도의 Frame을 잡고, 스펙트럼의 왜곡을 방지하기 위해 Windowing을 한다. 이때 windowing 함수는 보통 Hamming Window를 사용한다(식 3).

$$w(n)=0.54-0.64\cos\left(\frac{2}{N-1}\right), \quad 0\le n\le N-1 \qquad \text{(식 3)}$$

각각의 frame은 (식 4)에 의해 Autocorrelation을 구하고 Durbin의 algorithm에 의해 LPC 계수를 구한다(식 5).

$$r(m)=\sum_{n=0}^{N-1-m}\tilde{x_n}(n)\tilde{x_n}(m+m), \quad o\le m\le p \qquad \text{(식 4)}$$

$$E^{(0)}=r^{(0)}$$

$$k_i = \frac{[\ r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(|i-j|)]}{E^{(i-1)}}$$

$$a_i^{(i)} = k_i$$
$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}$$
$$E^{(i)} = (1-k_i^2)E^{(i-1)}$$

a_m= LPC coefficients = $a_m^{(p)}$, $1 \le m \le p$
k_m= PARCOR coefficients
$E^{(m)}$= Error for a predictor of order m (식 5)

구한 LPC 계수는 (식 6)에 의해 cepstrum 계수로 변환된다.

$$c_m = \begin{cases} \ln \delta_0^2, & n=0 \\ a_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, & n>0 \end{cases} \quad \text{(식 6)}$$

위 식으로 구해진 cepstrum 계수는 low quefrency 에 그 특성이 집중되어 있으므로 적절한 lifter가 필요하고, (식 7)를 사용한다.

$$l(n) = \begin{cases} 1 + \frac{L}{2} \sin(\frac{\pi n}{L}), & n=0,1,\ldots,L \\ 0, & other\ n \end{cases} \quad \text{(식 7)}$$

3. Frequency Sensitive Competitive Learning 신경회로망

FSCL 신경회로망은 단순 경쟁학습의 Dead 뉴런 문제를 보안하기위해 제안된 학습 알고리듬이며[2][3], 승자가 되는 빈도가 높은 뉴런에 대하여 양심(Conscience)을 부여하여 타 뉴런이 승자가 될 비율을 높여주는 것을 기본개념으로 한다. FSCL의 학습 알고리듬은 다음과 같다.

$$u_i = \begin{cases} 1, & \text{if } i=c \text{ such that } \gamma_c \|\vec{x} - \vec{w_c}\|^2 = \min_j \gamma_j \|\vec{x} - \vec{w_j}\|^2 \\ 0, & \text{otherwise} \end{cases}$$

$$\nabla \vec{w_i} = \begin{cases} \alpha(\vec{x} - \vec{w_i}), & \text{if } u_i = 1, \\ 0, & \text{otherwise.} \end{cases}$$

$\vec{u}$: FSCL출력
$\vec{x}$: 입력패턴
$\vec{w}$: 가중치 (식 8)
γ : $n_j / \sum_{i=1}^{k} n_i$
n_i : 뉴런 i 가 승자인 빈도 수

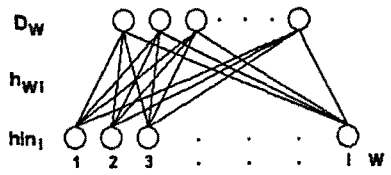
여기서 $\alpha(t)$는 학습률(Learning Rate)이며 초기치 α_0를 갖고 시간이 경과함에 따라 단조감소 하는 함수를 사용한다.

$$\alpha(t) = \alpha_0 (1 - \frac{t}{Number\ of\ Iteration}) \quad \text{(식 9)}$$

4. 음성 후처리부

30mS 정도의 frame 단위로 해석함으로 한 단어는 약 40-50개 frame을 갖는다. 즉 한 단어에 대하여 FSCL 신경회로망은 frame별로 순차적인 출력을 한다. 그런데 한 단어는 자음 모음등 벡터적으로 상이한 패턴으로 구성되어 있으므로 Map을 작성하기가 힘들다. 그러므로 여러개의 출력값들을 바탕으로 한개의 단어를 선택하는 기준이 필요하다.

후처리부의 구조는 (그림 2)와 같다. hin_i는 입력단어에 대한 출력뉴런 i의 승자뉴런인 횟수이고, D_w는 입력단어가 각 인식단어 w에 속하는 정도를 말한다.
FSCL신경회로망의 학습후, 인식단어 w에대한 출력뉴런 i가



(그림 2) 후처리부 구조

승자뉴런인 횟수 h_{wi}를 구하여 FSCL의 출력층과 인식단어 w와의 Weight를 다음과같이 정의하고.

$$h'_{wi} = \frac{h_{wi}}{\sqrt{\sum_{j=1}^{n} h_{wj}^2}} \quad \text{(식 10)}$$

D_w는 다음식에 의해 구해진다.

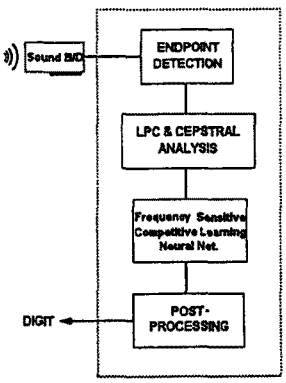
$$D_w = \sum_{j=1}^{i} h'_{wj} hin_j \quad \text{(식 11)}$$

결국 인식된 단어는 D_w가 가장 큰 w^*를 선택한다.

$$w^* = \arg\max(D_1, D_2, D_3, \ldots, D_w) \quad \text{(식 12)}$$

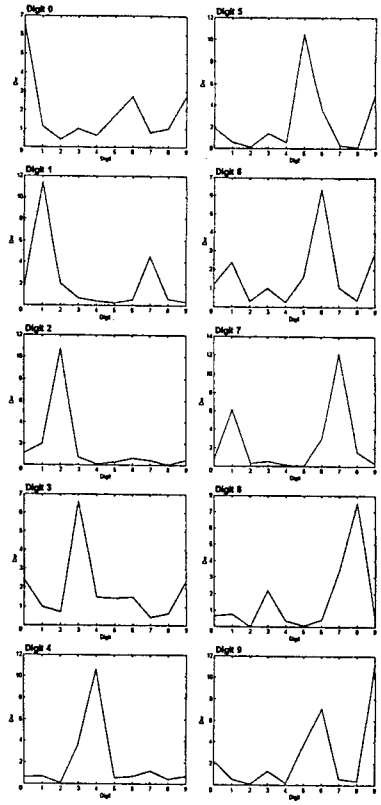
5 실험결과 및 고찰

본 논문에서 구현한 인식 시스템은 (그림 3)와 같다. 음성신호는 컴퓨터를 사용하는 실험실 환경에서 8bit resolution에 11 *kHz* sampling 하였다. 그리고 한 frame은 300samples이며, 100samples씩 shift한다.



(그림 3) 인식 시스템 블럭도

끝점검출(Endpoint Detection)은 시간 영역의 파라메터인 energy와 zero-crossing rate을 이용한 Rabiner와 Sambur의 algorithm을 사용하였다 [6]. 끝점검출후에 Durbin의 Algorithm으로 10차 LPC 계수를 구하고, 그로부터 Cepstrum 계수를 구한다. 이때 구한 10개의 Cepstrum 계수는 특징 벡터로 써 FSCL 신경회로망의 입력으로 사용된다. FSCL신경회로망의 학습패턴은 '영' '일' '이' '삼' '사' '오' '육' '칠' '팔' '구' 열개의 한국어 숫자음이며, 각 단어를 10회씩 발음한 100개 패턴, 4149개 frame,을 사용하였다. 이때 출력뉴런은 100 개 이고, (식 9)의 α_0 는 0.3, 학습은 각 frame당 70번의 Iterations을 수행하였다.
테스트 패턴은 학습하지 않은 음성을 각 단어당 10개씩 100개를 사용하였다. (그림 4)는 입력 음성에대한 D_w 의 예이다.



(그림 4) 입력음성이 각 숫자음에
속하는 정도 D_w

인식결과는 학습한 패턴에 대하여 98/100이며, 학습하지 않은 패턴은 96/100이다. 결과는 (표 1), (표 2) 이다.

발음한 단어 \ 인식한 단어	영	일	이	삼	사	오	육	칠	팔	구	인식률
영	10										10/10
일		9						1			9/10
이			10								10/10
삼				10							10/10
사					10						10/10
오						10					10/10
육							10				10/10
칠								10			10/10
팔				1					9		9/10
구										10	10/10
											98/100

(표 1) 학습한 패턴의 인식 결과

발음한 단어 \ 인식한 단어	영	일	이	삼	사	오	육	칠	팔	구	인식률
영	10										10/10
일		9						1			9/10
이			10								10/10
삼				10							10/10
사					10						10/10
오						10					10/10
육							10				10/10
칠		1						9			9/10
팔				2					8		8/10
구										10	10/10
											96/100

(표 2) 학습 하지않은 패턴의 인식 결과

6 결론

본 논문은 음성신호의 Cepstrum계수를 추출하여 FSCL 신경회로망을 사용하고 후처리를 추가하여 음성인식 시스템을 구현하였다. 인식시스템의 인식결과는 학습데이타에 대하여 98/100, 학습하지 않은 데이타에 대하여 96/100의 인식정도를 나타냈다. 자동 다이얼 전화기등에 본 연구의 결과가 실용화 될 수 있기 위해서는 향후에 화자독립 인식 시스템의 구현이 요구된다.

참고문헌

[1] J.R. Deller Jr, J.G. Proakis, J.H.L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, 1993.

[2] L.Xu, A Krzyzak, "Rival Penalized Competitive Learning for Clustering Analysis, RBF Net, and Curve Detection," IEEE Trans. on Neural Net. Vol. 4, NO.4, July 1993.

[3] S. C. Ahalt, A. K. Krishnamurty, P. Chen, and D. E. Melton, "Competitive learning algorithms for vector quantization," Neural Net. Vol. 3,1990.

[4] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall,1978.

[5] L.R. Rabiner and B.H. Jwang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[6] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell Syst. Tech. J. Vol. 54, No. 2,February 1975.

[7]김기석, 임은진, 황희용, "음성인식 신경망을 위한 음성 파라미터들의 성능 비교", 한국 음성학회지 11권 3호,1992.

[8] 길경하, 이천우, 박인정, "신경회로망을 이용한 음성계산기의 구현", 인공지능, 신경망 및 퍼지시스템 학술대회 논문집, 1993.

[9] 김동국, 정창균, 정 홍, "TDNN신경회로망을 이용한 한국어 음소인식", 음성통신 및 신호처리 Workshop, 1990.

[10] 최용석, 김장복 "Fuzzy 이론을 이용한 음성인식에 대한 연구", 석사학위논문, 1991.