인지 선형 예측 분석에 의한 음성 인식 방법

°김현철*, 송도선**, 김석동*

* 호서대학교 전자계산학과

** 중경공업전문대 전자계산기과

The Speech Recognition Method by Perceptual Linear Predictive Analysis

°Hyeon-chul Kim*, Do-sun Song**, Suk-dong Kim*

* Dept. of Computer Science in Hoseo University

** Dept. of Computer Engineering in Joongkyoung College

ABSTRACT

This paper proposes an algorithm for machine recognition of phonemes in continuous speech. The proposed algorithm is static strategy neural network. The algorithm uses, at the stage of train ng neuron, features such as PARCOR coefficient and auditory-like perceptual liner prediction(PLP). These features are extracted from speech samples selected by a sliding 25.6msec windows with s sliding gap being 3 msec long, then interleaved and summed up to 7 sets of parameters covering 171 msec worth of speech for use of neural inputs. Perfomances are compared when either PARCOR or auditory-like PLP is included in the feture set.

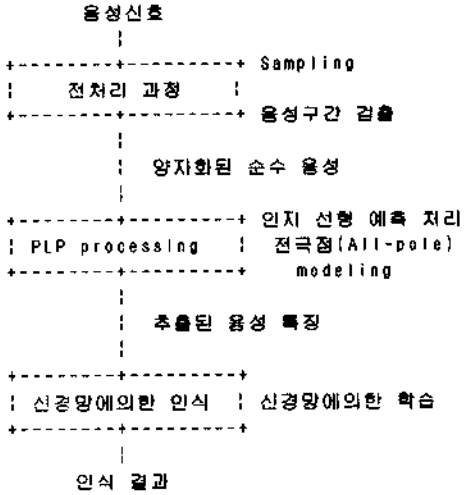
1 서론

음성은 사람들 사이의 대화에서 가장 자연스러운 정보전달수단이다. 만일 기계가 사람의 음성을 정확하고 신속하게 인식할 수 있다면 여러분야에서 유용하게 이용될 수 있을 것이다. 음성을 자동적으로 표기할 수 있고, 사람과 기계사이의 간단한 대화가 가능하며, 청각기능을 상실한 사람들에게 도움을 줄수가 있을 것이다. 그러나 아직까지 오늘날의 음성인식기로는 다양한 사람의 연속적인 음성을 인식하기에는 해결해야할 많은 문제점을 갖고있다. 사람에 따라 음성은 특유의 성질을 갖고 있으며, 시간에 따라 수시로 변화하는 것이 대표적인 문제로 볼 수 있다. 예를 들면, 어떤 사람은 보통 보다 빠르고/늦게 발음하거나, 크고/작게 말을 하며, 심지어는 같은 사람이라 해도 똑같은 단어를 동일한 형태로 발음하지 않는다. 이때 더구나 음성인식에서 올바르게 인식이 되었는지 결정하기도 쉽지가 않다. 일반적으로 사람의 경우 듣는 순간 5만개의 단어와 순간적으로 비교해가면서 음성학적이나 구문적이나 의미적인 다양한 분석을 행하면서 인식하고 있으나, 아직 음성인식기는 단순한 형태의 분석에 의해 인식되고 있는 실정이다. 음성언어는 생성되자 마자 바로 소멸되어 보조장비를 사용하지 않는한 그 전파범위가 지극히 한정적인 특성을 가지고 있으나 인간의 사고에 비해 생성되는데 걸리는 시간이 본 연구에서는 이러한 언어로 부터 화자 독립적인 특징을 구하고 음소간의 관계를 규명하여 음소단위로 인식할 수 있는 방법을 제시하고자 한다. 음성인식에 대한 연구는 음성의 전송및 합성 기술과 더불어 같은 역사적 배경을 가지고 발전되어 왔으며 그발전은 디지털 신호 처리 기술의 발달에 힘입은 바가 크다. 1952년 벨 연구소에서 최초로 formant 주파수를 이용하여 숫자음을 인식하는데 부분적으로 성공한 이후 간간히 연구결과가 소개되다가 1970년대 초반에 미국에서 행해진 DARPA프로젝트에 음성 이해에 대한 연구가 포함되면서 본격적인 연구가 시작되었다. 70년대 중반에는 일반적인 문제해결 알고리즘에 대해 상대적으로 특수한 지식의 중요성이 알려지면서 음성 그자체와 청각기관의 구조및 기능에 대한 연구로부터 시작되는 상향식 접근방법을 택하는 계기가 되었다. 벨 연구소의 Rabiner, IBM의 Jelinek, CMU의 Waibel, MIT의 Jue, Klatt, Lippman등의 연구자들이 전통적인 음성인식방법인 DTW (Dynamic Time Warping), HMM (Hidden Markov Model), VQ(Vector Quantization)등을 이용하여 연구를 행하였고 1984년 이후에 Kohonen, Lippmann, Anderson, Lang, Waibel, Sawai, Miyajake등이 인공신경망을 이용하여 음성인식에 적용하여 많은 성과를 얻고있다. 신경망이외에 퍼지이론을 도입한 연구도 있었고, HMM의 Viterbi 알고리즘을 신경망에 결합한 복합 기술연구도 발표된바 가 있다. 현재까지 음소 인식에 대한 연구로는 Kohonen의 Self-organizing feture map과 시간 지연 신경회로망 (Time-delay neural network)이 있다. Kohonen의 방법은 음소를 인식하

기 위한 학습방법이 자율학습으로 비슷한 음소끼리 묶어 분류하는 방법을 사용하였고, TDNN은 여러층을 갖는 역전파(Back Propagation)모델을 이용하여 각각의 연속적인 층으로 오랜시간에 대하여 음성을 적용하는 방법으로 가장 위층에서 시간에 불변하는 특성을 갖는 음소를 인식하는 방법으로 음소인식 기술중에 가장 성공한 경우이다. 그러나 이방법에는 방대한 신경망의 조직이 필요하고 학습시간이 매우 길어 소규모 컴퓨터로는 학습하기 어려운것이 문제점이다.

2 음성인식 방법

본 연구에서 사용할 전체적인 방법은 아래와 같다.



2.1 전처리과정

PC486하에서 DT2801A 보드를 사용하여 10kHz 12bit 양자화하였다. 데이타의 입력을 위하여 프로그램은 단구간 에너지를 이용하여 음성의 시작점을 자동으로 검출하도록 작성되었고 음성의 시작점으로부터 1.6초간 입력을 받을 수 있다. 또한 입력된 음성을 적당한 크기로 나누어 텍스트 파일로 저장할 수 있다.

2.2 인지선형예측 처리

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

▶ PLP 캡스트럼 계수(cep) : 제 7 차 인지 선형 예측 계수를 이용한 캡스트럼 계수이다. PLP를 이용한 캡스트럼 계수는 PLP 계수로부터 다음과 같이 얻을 수 있다.

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

PLP 계수를 포함한 LP계수는 차수가 올라갈수록 작은 크기의 값을 갖는다. 본 논문에서는 모든 계수가 비슷한 범위를 갖도록 차수를 곱하여 표현하였다. 본 논문에서는 모든 계수가 비슷한 범위를 갖도록 차수를 곱하여 표현하였다. 즉,

$$Ci = Ci \cdot (i+1) \qquad i=1,2,\cdots,p \quad (2\text{-}2)$$

▶ 단구간 에너지 : 25.6ms의 음성 신호에 해밍 윈도우 처리를 한 제곱합이며, PLP 처리과정에서 상당부분 왜곡된다.

▶ 단구간 영교차율 : 단위 시간에 0를 지나는 횟수를 말하며 256 표본의 음성신호중 0를 지나는 횟수를 계수하여 사용하였다.

2.3 학습을 위한 학습패턴의 구성

학습패턴은 인식기의 학습을 위한 목표값이며 올바른 목표치를 사용하여야 올바르게 학습을 시킬 수 있다. 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. 그러므로 다음과 같은 방법으로 올바른 목표치를 정하고 음소의 경계에 대하여 일정구간 학습을 금지 시킴으로써 목적을 달성할 수 있다. 이와같은 구성은 음소경계에 대한 애매한 출력, 즉 경계 양쪽 음소의 특징을 모두 포함하는 출력을 내도록 인식기를 유도할 것이며, 이것이 바람직한 출력이다. 그림[2-1]은 목표치를 설정하는 방법을 나타낸 것이다. 유영으로 나타낸 부분은 주위 음소와 학습에 영향을 미치지만 자신은 학습되지 않는다. 각각의 음소의 길이는 단어에 따라 달라지며 음영으로 표시된 부분은 20ms로 임의로 설정하여 학습에서 제외시켰다.

그림 2-1 학습을 위한 목표치의 설정

학습을 위한 목표치의 입력은 파형과 스펙트럼이 동시에 표시된 화면상에서 행해진다. 화면을 주시하면서 키보드를 이용하여 세로막대 커서를 이동시킨 후 순차적으로 목표값을 입력하면 프로그램은 좌표와 목표값을 입력받아 해당하는 위치의 프레임들에 대하여 입력값과 목표치를 파일에 기록한다. 연속음에 나타나는 음소는 전후의 다른 음소와의 조음 현상에 의한 변화가 심하다. 현재의 음성 조각을 인식하기 위해서는 인접한 음소를 같이 참조해 보 것. 다친해유, 더더욱 사림의 그것과 유사해진다. 보통은 이러한 변화를 학습하기 위하여 시간지연신경망을 사용하지만 본 논문에서는 고정 신경망을 사용하고, 입력단의 구성을 달리하여

시간 변화를 학습하였다. 3ms의 음성 프레임을 인식하기 위하여 특징들의 집합을 구성한다. 집합은 총 63개의 입력으로 구성되며, 각각의 요소에 대하여 [그림2-2]에 도시하였다.

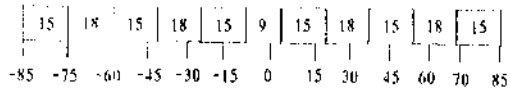


그림 2-2 입력벡터의 요소 구성

사선은 PLP가 평균되어지는 구간의 길이를 나타내며, 점선은 사용되지 않는 구간의 길이이다. 현재의 프레임에 대하여 인접한 3개의 프레임을 평균하였으며, 나머지는 인접한 다섯개의 프레임을 구하여 평균을 취하였다. 각각의 프레임은 9개의 데이타로 구성되어 있다. 7개의 PLP 계수와 1개의 단구간 에너지 그리고 1개의 영교차율이 그것이다.

2.4 학습 및 인식

음어한 국어소 클로리가은 음성의 입력을 한국어의 음소를 가로소 보내줄 뿐, 다 한음 들력으로 내보낸다. 이러한 음소들은 'ㄱ', 'ㄴ', 'ㄷ', 'ㄹ', 'ㅁ', 'ㅂ', 'ㅅ', 'ㅇ', 'ㅈ', 'ㅉ', 'ㅊ', 'ㅋ', 'ㅌ', 'ㅍ', 'ㅎ'의 19개의 자음과, 'ㅏ' 'ㅑ', 'ㅓ', 'ㅕ', 'ㅗ', 'ㅛ', 'ㅜ', 'ㅠ', '..', '.', '총', 'ㅐ' 등의 12개의 모음과 'space' 1개의 총 32개 호 대하여, 특 3m 호에 25.6m 대 여하영역 특호에 대하여 특을 실한다(실한 . 다시을 실시한다.

3. 실험및 검토

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

음성의 채집은 PC486 기종에서 DT2801 보드를 이용하여 10 kHz로 12bit 양자화하여 사용하였다. 데이타의 입력은 실내의 비교적 소음이 적은곳에서 음성의 시작점을 검출하여 데이타를 입력한다. 데이터의 처리를 위하여 PC용 UNIX인 Linux 환경하에서 GNU C++를 사용하여 프로그램이 작성되었다. UNIX 기반 OS인 Linux를 사용하는 잇점은, 일반적으로 신호처리가 대규모의 메모리를 요구하는데, 거의 무제한의 메모리(Virtual Memory)를 사용할 수 있다는 점과 고해상도의 그래픽을 사용할 수 있다는 데에 있다. 또한 빠른 처리속도와 MS-DOS와의 데이타 호환성도 무시할 수 없는 잇점이다. 그래픽의 처리를 위하여 함께 제공되는 'vga library'를 사용하였고, 논문 전반에 걸쳐 여기에서 처리한 결과를 도시하였다.

마지막 단계인 인식기는 마이크로소프트 윈도우즈 3.1환경하에서 볼랜드C++3.1을 사용하여 객체지향 기법으로 작성하였다. 메타와 행열등의 처리를 위하여 클래스를 정의하였으며, 필요한 연산자들도 오버로딩함수로 정의되어 프로그램의 가독성을 증가시키고 기능을 보강화 하였다. 한번에 설정할 수 있는 배열의 제한으로 인하여 입력 데이타는 선형 리스트를 구성하여 처리하였다. 실험대상 단어는 한사람이 발음한 것 중 각각 한개씩을 사용하여 학습하였다. 이제에 나타낸 표는 학습된 인공신경망이 학습된 사람을 포함하여 5명이 발음한 음성을 인식한 결과이다. 학습을 위하여 한사람이 발음한 음성중 각각 1개씩을 사용하였으며 나머지와 다른 사람이 발음한 음성을 사용하여 아래와 같은 인식률을 표를 얻었다. 표의 각각의 칸은 음소의 개수와 인식된 갯수를 나타내며 오른쪽에 각 음소에 대한 평균인식률과 아래쪽에 각 화자에 대한 평균 인식률을 나타내었다.

	A	B	C	D	E	평 균
ㅏ	5/5	5/5	5/5	5/5	5/5	100
ㄴ	4/5	3/5	4/5	1/5	3/5	64
ㅣ	4/5	0/5	0/5	5/5	5/5	56
ㅇ	3/5	1/5	2/5	5/5	4/5	60
ㅏ	5/5	5/5	5/5	5/5	5/5	100
ㅅ	5/5	5/5	5/5	5/5	5/5	100
ㅣ	4/5	4/5	3/5	1/5	0/5	48
ㅁ	5/5	5/5	5/5	5/5	2/5	88
ㅣ	5/5	5/5	5/5	5/5	5/5	100
SP	5/5	5/5	4/5	5/5	5/5	100
ㅁ	3/5	0/5	5/5	3/5	3/5	56
ㅏ	5/5	5/5	5/5	5/5	5/5	100
평 균	96.7	71.7	80.0	81.7	78.3	81.7

	A	B	C	D	E	평 균
ㅇ	5/5	5/5	5/5	5/5	5/5	100
ㅏ	5/5	4/5	5/5	5/5	5/5	96
ㄴ	4/5	4/5	4/5	4/5	5/5	84
ㅅ	5/5	3/5	2/5	4/5	2/5	64
ㅏ	5/5	5/5	5/5	5/5	5/5	100
ㄴ	5/5	5/5	5/5	4/5	5/5	96
ㅣ	5/5	5/5	5/5	5/5	5/5	100
ㅎ	5/5	5/5	4/5	5/5	5/5	96
ㅏ	5/5	5/5	5/5	5/5	5/5	100
평 균	97.8	88.9	88.9	93.3	93.3	92.9

[표4-1]과 [표4-2]에 나타낸 것과 같이 제안된 시스템의 화자 의존도를 알 수 있다. 화자 A의 음성을 사용하여 학습하였기 때문에 화자 A의 음성에 대한 인식률이 비교적 높게 나타났다. 각각 한개씩만을 학습에 참여시켰기 때문에 학습한 화자마저도 완전한 출력을 내지는 못하였으나 학습하는 단어의 수가 증가할수록 시스템의 인식률은 높아질 것이다. 학습에 참여하지 않은 화자 음성에 대한 출력도 어느정도 원하는 값에 접근하였으며 이것으로써 제안된 시스템이 화자 독립적인 특성을 가진다고 볼 수

있다.

4. 결론

단일 화자 학습 시스템에서의 음소의 인식률은 모음보다는 음소의 연음현상이 심한 자음 부분에서 비교적 저조한 인식률을 나타내었다. 이것은 사람에 따라 다른 변화를 가지는 음성을 단일화자 학습 시스템에서는 충분히 학습하지 못한 결과라고 판단된다. 또한 소규모의 음성을 대상으로 한 실험이기 때문에 인식률이 우수하게 나타났다고 생각되며, 더 많은 변화를 수반하는 많은 음성을 대상으로 실험한다면 더 좋은 결과를 낼 수도 있으나 반대의 결과도 예상된다. 이것의 해결방안으로는 제안된 시스템으로 해결되지 않는 부분에 대하여 다른 방법으로 인식을 시도할 수 있을 것이다. 또한 많은 음성을 대상으로 학습을 하는 경우 계층구조를 가지는 인식시스템을 구성하는 것도 바람직 하리라고 판단된다.

5. 참고문헌

[1] T.W. Parsons, *Voice And Speech Processing*, McGraw-Hill Book Company. 1986

[2] R.D. Pearoke and D.H.Graf, "An Introduction to Speech and Speaker Recognition," Computer, vol. 23, no. 8, August 1993.

[3] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Engelwood Cliffs, N.J., Prentice-Hall, 1978.

[4] R.P. Lipmann, "Review of Neural Networks for Speech Recognition," *Readings in Speech Recognition*, Morgan Kufmann, Inc., 1992.

[5] H.Sakoe and S.Chiba, "Dynamic Programing Optimazition for Spoken Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing. Vol. ASSP-26, pp. 43-49, Feb. 1978.

[6] Lee, Kai Fu and H.W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," Proceedings IEEE ICASSP. pp.123-126, 1988.

[7] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang, "Phoneme Recognition using Time-Delay Neural Network" IEEE Trans. Vol. ASSP-37, No. 8, Aug. 1989.

[8] A.Waibel T.Hanazawa, G.Hinton, K.Shikano, and K.Lang : "Phoneme Recognition: Neural Networks vs. Hidden Markov Models," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY, April 1988

[9] R.P. Lipmman, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, Vol.4, No. 2, pp. 4-22, April 1987.

[10] D.J.Burr : "Experiments on Neural Net Recognition of Spoken and Written Text," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 36, No.7, pp.11622-1168, July 1988.

[11] 김석동, 이행세, "신경망을 이용한 우리말 인식에 관한 연구," 한국음향학회지 제11권 3호, 1992.

[12] G.M. White and R.B. Neely, "Speech Recognition Experiment with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoust., Speech and signal Processing, Vol. ASSP, April 1976, 183-188