論文95-32B-8-8

단어사전과 다층 퍼셉트론을 이용한 고립단어 인식 알고리듬

(Isolated Word Recognition Algorithm Using Lexicon and Multi-layer Perceptron)

李 基 熙 *, 林 寅 七 **

(Kee Hee Lee and In Chil Lim)

요 약

최근 음성신호를 신뢰성있게 인식하기 위한 다양한 기법이 개발되고 있다. 우수한 패턴인식 특성을 갖는 다층 퍼셉트론(MLP)은 음성인식 분야에서 가장 용도가 넓은 네트워크 중의 하나이다. 이 논문은 MLP와 단어사전을 이용한 자동 음성인식 시스템에 대해 기술한다. 이 시스템에서는 단어사전의 단어를 MLP의 출력값에 정합시키는 네트워크 탐색 알고리듬에 의해 음성 인식이 수행된다. 또, 음소의 지속시간 정보를 고려하여 기존의 CHMM(Continuous HMM)에 필적하는 성능을 가지는 인식 알고리듬도 제시한다. 제안한 시스템의 성능은 9명의 화자가 발음한 26개 어휘의 데이타 베이스로부터 평가된다. 실험결과는 제안한 알고리듬의 오인식률이 CHMM의 12.6%보다 5.3% 낮은 7.3%임을 보인다.

Abstract

Over the past few years, a wide variety of techniques have been developed which make a reliable recognition of speech signal. Multi-layer perceptron(MLP) which has excellent pattern recognition properties is one of the most versatile networks in the area of speech recognition. This paper describes an automatic speech recognition system which use both MLP and lexicon. In this system, the recognition is performed by a network search algorithm which matches words in lexicon to MLP output scores. We also suggest a recognition algorithm which incorperate durational information of each phone, whose performance is comparable to that of conventional continuous HMM(CHMM). Performance of the system is evaluated on the database of 26 vocabulary size from 9 speakers. The experimental results show that the proposed algorithm achieves error rate of 7.3% which is 5.3% lower rate than 12.6% of CHMM.

* 正會員, 大宥工業專門大學 事務自動化科
(Dept. of Office Auto., Daeyeu Tech. Coll.)
** 正會員, 漢陽大學校 電子工學科
(Dept. of Elec. Eng., Hanyang Univ.)
接受日字:1995年 5月 2日, 수정완료일:1995年 7月 31日

I. 서 론

음성은 사회의 정보화가 급속히 진전되면서 인간과 기계간의 정보전달을 위한 매개체로서의 역할과 그 이용이 더욱 증대되고 있다. 이를 위한 하나의 방법으로 인간이 발성한 음성을 컴퓨터가 인식하는 음성인식에 관한 연구가 꾸준히 진행되고 있다. 일반적으로 많이

사용하는 인식 알고리듬으로는 DTW(Dynamic Time Warping)[1]와 HMM(Hidden Markov Model)[2-7]을 들 수 있다. DTW는 기준 템플레이트(template)와 입력음성을 시간적으로 정합하여 유사도(likelihood)가 높은 기준 템플레이트의 음성으로 인식하는 방법이다. 한편, HMM을 이용한 방법은 기준 HMM에서 입력 음성이 관측될 확률을 구하여 가장 높은 확률을 가지는 HMM의 음성으로 인식한다.

최근에는 신경망(neural network)[8-12] 이론을 음성인식에 도입하여 보다 우수한 성능을 가지는 음성인식기가 연구되고 있다. 신경망은 인간의 뇌세포에 해당하는 처리요소들이 신경에 대응하는 정보채널을 통해 연결된 망으로, 처리요소들이 병렬로 구성되어 있어서 대량의 복잡한 데이타를 분산 처리할 수 있다. 또한 학습을 통해 음성의 특징을 찾아서 적응할 수 있는 능력이 있으므로 우수한 패턴 인식특성을 가진다. 음성인식에 사용되는 대표적인 신경망으로는 다층 퍼셉트론(MLP:Multi-Layer Perceptron)[10]과 시간지연 신경망(TDNN:Time-Delay Neural Network)[9,12,15]을 들 수 있다.

본 논문에서는 MLP의 출력값(output score)과 네트워크 탐색 알고리듬(searching algorithm)을 이용한 단어인식 시스템을 구성하였다. MLP는 음성신호를 음소단위로 인식하여 각 음소에 대응되는 출력층의 값을 발생시킨다. 탐색 알고리듬은 단어사전(lexicon)의 단어를 MLP의 출력값에 정합하여 음성을 인식하게 된다. 음소의 수는 단어나 음절의 수에 비해 매우 적으므로, 대규모 어휘 인식에는 음소단위의 음성인식이 효과적이다. 그러나 음소는 그 특성상 문맥이나 화자의 발성습관 또는 조음효과(coarticulation effect) 등의 영향으로 다양한 변화를 보이므로 음소를 신뢰성있게 인식하는 것은 어렵다. MLP는 학습을 통해 음소의 특징을 찾아 적응하는 능력이 있으므로 음소를 어느 정도 분류 인식할 수 있다. 탐색 알고리듬은 단어사전을 이용하여 최적의 음소열을 찾으므로 MLP의 인식 오류가 발생하더라도 이에 대처할 수 있다. 이 시스템에서는 기준 모델이나 템플레이트를 이용하지 않고 단어사전의 단어를 MLP의 출력값에 효과적으로 정합함으로써 기존의 음성 인식기에서 필요한 메모리 용량과 계산량을 크게 줄일 수 있다. 제안한 인식 시스템의 성능은 단어인식 실험을 통해 확인하였다.

Ⅱ. 단어인식 시스템

1. 인식 시스템의 개요

단어음 인식 시스템은 그림 1과 같이 구성된다. 음성신호가 입력되면 전처리 과정을 거쳐 특징 파라미터를 추출한 후, 다층 퍼셉트론(MLP:Multi-Layer Perceptron)을 이용하여 매 프레임별 MLP의 출력값을 구한다. 인식 알고리듬은 MLP의 출력값, 최소 지속시간 및 단어사전을 이용하여 음소단위의 단어인식을 수행하게 된다.

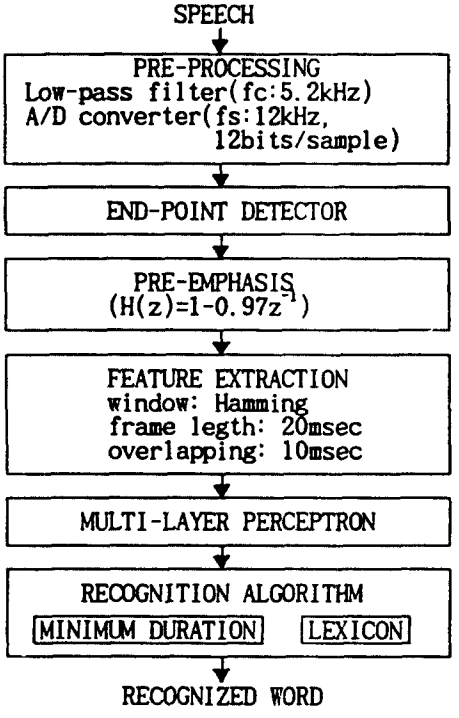

그림 1. 인식 시스템의 블럭도
Fig. 1. Block diagram of the recognition system.

2. 전처리 과정 및 특징 추출

입력된 음성신호를 차단주파수가 5.2kHz인 저역통과 필터에 통과시킨 후, 12kHz로 표본화하고, 12bit로 양자화한다. 전처리 과정에서는 전달함수가 $1-0.97z^{-1}$인 디지틀 필터를 통과시켜 고주파 성분을 강조한다. 끝점 검출기에서는 Rabiner와 Sambur가 제안한 방법[13]을 이용하여 음성의 시작점과 끝점을

검출한다. 다음으로 음성을 20msec 구간(240 samples) 프레임으로 분할하고, 각 프레임을 10msec 씩 중첩하여 진행하면서 해밍윈도우 함수를 이용, 특징 벡터를 추출한다. 음소의 특징을 표현하는 파라미터로는 선형예측(linear prediction) 분석에 의해 구한 16차 LPC 켑스트럼과 로그에너지를 사용한다. 다층 퍼셉트론은 현재의 프레임과 전/후 한 프레임씩의 특징 파라미터를 입력으로 하여 각 음소별 출력값을 계산한다.

3. 다층 퍼셉트론(MLP)

인식 대상 단어는 "영"부터 "구"까지와 "공"을 포함한 숫자음 11가지와 대도시명 15가지로서 모두 26개이며, 이에 포함된 음소는 19종류이다. 이를 하나의 MLP으로 학습할 경우, MLP의 학습에 소요되는 시간이 늘어나고 학습에 필요한 데이타의 수도 더욱 많아지게 된다. 본 논문에서는 그림 2와 같이 전체 음소를 모음, 비음 및 무성음 등 3개의 모듈로 나누어 각 모듈별로 다층 퍼셉트론을 학습한다[14]. MLP의 입력은 현재 프레임과 전/후 프레임의 특징벡터인 51개의 파라미터로서, 은닉층 1 및 2를 거쳐 출력층으로 들어가게 된다.

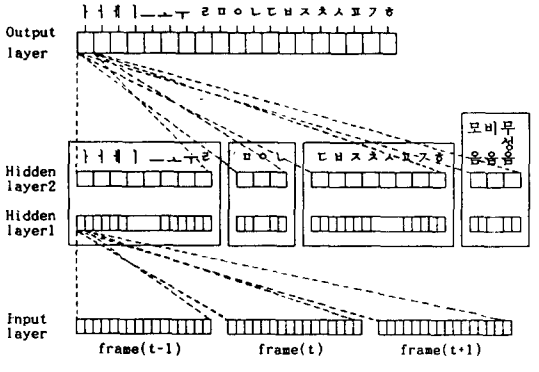

그림 2. 다층 퍼셉트론의 구조
Fig. 2. Structure of multi-layer perceptron.

그림 2에서 맨 우측의 MLP는 입력 특징벡터를 모듈별로 분류하기 위한 모듈분류 MLP로서, 이의 출력값도 출력층으로 들어가게 된다. 모음, 비음 및 무성음은 음성적인 특징이 현저히 다르므로 모듈분류 MLP는 입력음성을 3가지중 하나의 모듈로 분류한다. 이때 모듈 분류망의 은닉층 2의 값은 자기 음소 모듈에 대해서는 1로, 다른 음소 모듈에 대해서는 0으로 각각 학습된다. 그림의 맨 위의 출력층에서는 각 모듈별 출력값과 모듈분류 MLP의 출력값으로 부터 출력노드의 값을 계산하게 된다. 출력층 및 각 모듈의 MLP에 연결된 가중치는 오차 역전파 알고리듬[11]을 이용하여 학습된다.

Ⅲ. 단어인식 알고리듬

이 절에서는 음소단위의 단어인식과 인식 알고리듬에 대해 기술한다.

1. 음소단위의 단어인식

고립단어 인식은 인식할 단어가 많아지면 각 단어별 기준 모델 또는 기준 템플레이트의 수가 늘어나므로 인식기의 메모리 용량 및 계산량이 많아지는 단점이 있다. 대규모의 어휘를 인식하기 위해서는 음성신호를 보다 작은 단위 즉, 음절, 반음절 또는 음소단위로 나누어 인식할 필요가 있다. 특히, 음소와 같은 미소한 음성적 단위로 인식할 경우, 음소의 종류가 대략 50~60개에 불과하므로 매우 효과적으로 대규모 어휘를 인식할 수 있다. 그러나 조음효과나 화자의 발성습관에 따른 불분명한 발음 등의 영향으로 음소인식 자체가 어려워지므로 음소를 인식단위로 하는 단어인식은 매우 어렵게 된다. 본 논문에서는 MLP와 단어사전을 이용한 음소단위의 단어인식 시스템을 구성하여 이의 성능을 확인하였다.

2. 네트워크 탐색 알고리듬

인식할 음소를 P(j)($1 \le j \le P$, P는 음소의 수)로, 시간(프레임) t에서 MLP의 j번째 노드의 출력값을 $O_t(j)$($1 \le t \le T$, T는 입력음성의 프레임 수)로 각각 표현하기로 하자. 미지의 입력음성은 전처리 과정과 MLP을 거쳐 $O_t(j)$로 변환한다. 한 예로서 음성 "부산"에 대한 파형과 $O_t(j)$의 값을 그림 3에서 보였다. 그림 3(b)에서 세로축은 음소를, 가로축은 시간을 각각 나타내며, 직사각형이 클수록 $O_t(j)$가 1에 가까움을 나타낸다. 음소 '우'부분에서는 음소 'ㄴ'의 값이 커지기도 하며, 음소 'ㅅ' 부분에서는 음소 'ㅈ'의 값도 큰 값이 되기도 한다. 그러나 $O_t(j)$의 큰 값만 따라가 보면 전체적으로 "부산"에 가까운 형태가 됨을 알 수 있다. 인식 알고리듬은 기본적으로 시간 t=1~T까지의 전체 프레임에서 모든 가능한 경로중 Ot(j)의 누적값이 최

대가 되는 경로에 포함된 음소열을 인식된 단어로 한다. 그림 3에서 Ot(j)가 최대값만 탐색하면 그림 4(a)와 같은 경로가 되어 이 경우의 인식된 음소열은 "ㅂ, 우, 으, ㄴ, ㅅ, ㅈ, ㅅ, 아, 으, ㅁ, ㄴ"와 같이 매우 복잡하게 된다. 반면, 그림 4(b)는 바람직한 경로를 나타낸 것으로, 비록 일부 구간의 출력값은 최대값이 아니지만 전체적인 면에서 타당한 음운 배열이 된다.

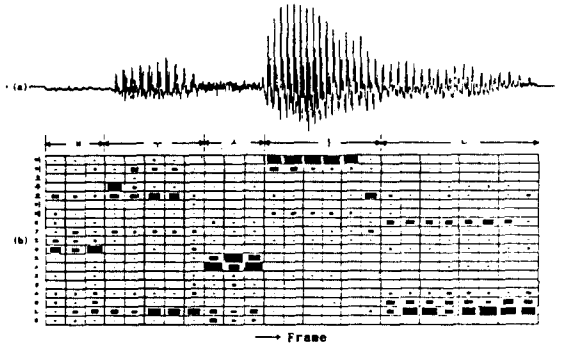

그림 3. (a) 단어 "부산"의 파형, (b) MLP의 출력값

Fig. 3. (a) Waveform of word "Busan", (b) Output scores of MLP.

이와 같이 음소단위의 단어인식이 잘 되지 못하는 원인으로는 1) 음소의 시작 부분이나 끝 부분에서의 불분명한 포만트 구조, 2) 인접한 음소간의 불분명한 경계점, 그리고 3) 조음효과 등을 들 수 있다. 이러한 문제점을 해결하는 한 방안으로 다음과 같이 단어사전을 이용할 수 있다. 단어사전은 인식 대상 단어의 음소를 순서대로 나열해 둔 것이다. V개의 단어를 가지는 단어사전에서 ν 번째

$$W^{(\nu)} = \{j_1, j_2, \cdots, j_n, \cdots, j_N\} \quad (1)$$

단어로 표시된다. 여기서 j_n은 음소의 번호이고, N은 단어 $W^{(\nu)}$의 음소 수이다. 그림 3(b)와 같은 격자형 $O_t(j)$의 배열에 단어 $W^{(\nu)}$를 정합시키는 과정은 다음과 같다. 먼저 $W^{(\nu)}$의 첫번째 음소 $P(j_1)$에 대해, $O_t(j_1)$의 누적값 $\delta_t(j_1)$을 구한다. 즉,

$$\delta_t^{(\nu)}(j_1) = \delta_{t-1}^{(\nu)}(j_1) + O_t(j_1) \quad (2)$$

여기서 $\delta_t^{(\nu)}(j_1)$은 음소 $P(j_1)$을 따라가며, 시간 t까지 $O_t(j_1)$을 누적한 것이다. $W^{(\nu)}$의 두번째 음소 $P(j_1)$에 대해서 시간 t까지의 누적값 $\delta_t^{(\nu)}(j_2)$는

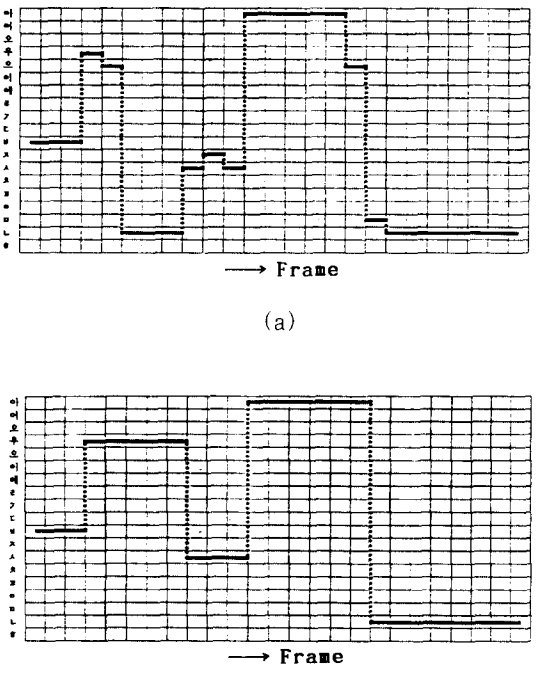

그림 4. 네트워크 탐색 알고리듬의 설명.
(a) 최대값에 따른 경로
(b) 바람직한 경로

Fig. 4. Illustration of the network search algorithm.
(a) path of maxinum output scores,
(b) Desired path.

$$\delta_t^{(\nu)}(j_2) = \max\left[\delta_t^{(\nu)}(j_1), \delta_{t-1}^{(\nu)}(j_2)\right] + O_t(j_2) \quad (3)$$

로 구한다. 즉, 시간 (t-1)에서 이전의 음소 $P(j_1)$의 누적값 $\delta_{t-1}^{(\nu)}(j_1)$과 현재의 음소 $P(j_2)$의 누적값 $\delta_{t-1}^{(\nu)}(j_2)$를 비교하여 큰 값으로 택한다. 그림 5에서 보인 것과 같이 시간 t에서 현재의 음소 $P(j_2)$로 천이하는 음소는 시간 (t-1)에서 음소 $P(j_1)$이 $P(j_2)$로만 제한된다. 이것은 MLP에서 음소인식이 잘못되더라도 강제적으로 정합시키기 위한 것이다.

위의 과정을 단어 $W^{(\nu)}$의 맨 마지막 음소 $P(j_N)$까지 수행하여 누적값 $\delta_T^{(\nu)}(j_N)$이 최대가 되는 단어를 인식된 W_{rec}으로 판정 한다. 즉,

$$\nu^* = \text{argmax}\ \delta_T^{(\nu)}(j_N) \quad (4)$$

$$W_{rec} = W^{\nu^*} \quad (5)$$

이상의 인식과정을 정리하면 다음과 같다.

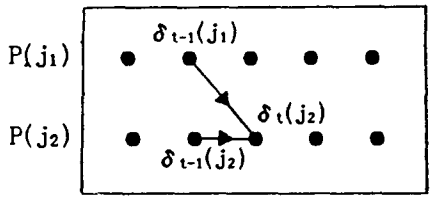

그림 5. 음소 천이의 설명
Fig. 5. Illustration of phone transition.

[알고리듬 1]

단계 1: 초기화

$\delta_0^{(\nu)}(j) = 0,\ 1 \le \nu \le V, 1 \le j \le P$

으로 둔다.

단계 2: 반복수행

a) $\nu = 1,2,\cdots,V$에 대해 b)~c)를 수행한다.

b) 단어 $W^{(\nu)}$의 첫번째 음소 $P(j_1)$에 대해

$\delta_t^{(\nu)}(j_1) = \delta_{t-1}^{(\nu)}(j_1) + O_t(j_1),\ 1 \le t \le T$

을 구한다.

c) 단어 $W^{(\nu)}$의 p번째 음소 $P(j_p)$에 대해

$\delta_t^{(\nu)}(j_p) = \max\left[\delta_{t-1}^{(\nu)}(j_{p-1}), \delta_{t-1}^{(\nu)}(j_p)\right] + O_t(j_p),$

$1 \le t \le T, 2 \le p \le N$

를 구한다.

단계 3: 인식된 단어 W_{rec}

$\nu^* = \arg\max \delta_T^{(\nu)}(j_N), 1 \le \nu \le V$

$W_{rec} = W^{\nu^*}$

이다.

3. 지속시간 정보를 고려한 네트워크 탐색 알고리듬

음소의 지속시간은 음소에 따라 다양한 변화를 보이고 있다. 대체로 자음은 짧게 모음은 길게 나타나며, 어떤 단어를 천천히 발음하면 자음보다는 모음의 길이가 늘어나게 된다. 또한, 화자의 발성습관이나 단어에 따라 지속시간은 다양한 분포를 보인다. 본 논문에서는 각 음소별 지속시간의 최소치를 고려하여 인식에 사용하였다. 지속시간의 최소치는 사람의 청각기관에 의해 음소로서 인지될 수 있는 최소한의 길이를 의미하며, 대체로 초성 자음은 짧게, 모음이나 종성 자음은 비교적 길게 나타난다.

인식 알고리듬에 최소 지속시간을 고려하려면 알고리듬 1의 단계 2 c) 에서 이전 음소 $P(j_{p-1})$의 지속시간이 이 음소의 최소 지속시간보다 짧을 경우, 현재 음소의 $P(j_p)$로 천이할 수 없도록 제한하면 된다. 각 음소 P(j)의 최소 지속시간 을 $P_{min}(j)$로, 단어 ν 의 음소 P(j)에서 현재의 지속시간을 $D_t^{(\nu)}(j)$로 두고 인식 알고리듬을 정리하면 다음과 같다.

[알고리듬 2]

단계 1 : 초기화

$\delta_0^{(\nu)}(j) = 0,\ \ 1 \le \nu \le V, 1 \le j \le P$

$D_0^{(\nu)}(j) = 0$

으로 둔다.

단계 2 : 반복수행

a) $\nu = 1,2,\cdots,V$에 대해 2 b)~2 c)를 수행한다.

b) 단어 $W^{(\nu)}$의 첫번째 음소 $P(j_1)$에서,

$t_1 = 1$

$\delta_t^{(\nu)}(j_1) = \delta_{t-1}^{(\nu)}(j_1) + O_t(j_1),\ 1 \le t \le T$

$D_t^{(\nu)}(j_1) = D_{t-1}^{(\nu)}(j_1) + 1$

을 구한다.

c) 단어 $W^{(\nu)}$의 $p = 2,3,\cdots,N$에 대해

$t_p = t_{p-1} + D_{\min}(j_{p-1})$

로 두고 $t = t_p, \cdots, T$까지

$\delta_{t-1}^{(\nu)}(j_{p-1}) > \delta_{t-1}^{(\nu)}(j_p)$ 이고 $D_{t-1}^{(\nu)}(j_{p-1}) \ge D_{\min}(j_{p-1})$이면

$\delta_t^{(\nu)}(j_p) = \delta_{t-1}^{(\nu)}(j_{p-1}) + O_t(j_p)$

$D_t^{(\nu)}(j_p) = 1$

로 두고, 그렇지 않으면

$\delta_t^{(\nu)}(j_p) = \delta_{t-1}^{(\nu)}(j_p) + O_t(j_p)$

$D_t^{(\nu)}(j_p) = D_{t-1}^{(\nu)}(j_p) + 1$

로 둔다.

단계 3 : 인식된 단어 W_{rec}

$\nu^* = \arg\max \delta_T^{(\nu)}(j_N),\ 1 \le \nu \le V$

$W_{rec} = W^{\nu^*}$

이다.

위의 알고리듬 단계 2 c)에서 t_p는 (p-1)번째 음소까지의 최소 지속시간의 합으로서, t <tp인 구간에서는 최소 지속시간보다 짧은 음소가 있게 되므로 이를 제거하기 위하여 t≥tp인 구간에서만 $\delta_t^{(i)}(j_p)$를 구하도록 하였다.

Ⅳ. 실험 및 결과

제안한 음성인식 시스템의 성능평가를 위해 단어음성에 대한 인식실험을 수행하였다.

1. 음성 데이타 채집과 MLP의 학습

인식실험에 사용된 음성 데이타는 표 1과 같이 11개의 숫자음과 15개의 도시명으로 이루어진 26개의 고립단어음을 남성화자 9명이 각각 20회 발음하여 얻어진 4680개로 구성된다. 음소 데이타는 26개의 단어음에 내포된 19개의 음소를 남성화자 9인이 20회씩 발음한 3420개의 음소 데이타로서 이중 1710개는 MLP의 학습에 이용하였으며, 나머지 1710개는 음소인식 실험에 사용하였다.

표 1. 음성 데이타
Table 1. Speech data.

숫 자 음	공 영 일 이 삼 사 오 육 칠 팔 구
도 시 명	서울 부산 대구 대전 인천 광주 강릉 청주 전주 제주 김천 성남 경주 포항 춘천

MLP는 그림 2에서 보인 것과 같이 유사 음소간의 변별력을 높이고 학습시간을 줄이기 위해 모음 모듈(ㅏ,ㅓ,ㅔ,ㅣ,ㅡ,ㅗ,ㅜ,ㄹ), 비음 모듈(ㅁ,ㅇ,ㄴ), 무성음 모듈(ㄷ,ㅂ,ㅈ,ㅊ,ㅅ,ㅍ,ㄱ,ㅎ)로 분류하여 각 모듈에 속한 음소간의 차이만 구분하도록 학습하게 하였다. 또 음소 'ㅖ'와 'ㅐ'는 구분이 어려우므로 'ㅔ'로 통일하였다.

2. 음소인식 실험

표 2는 9명의 화자가 발음한 1710개의 음소에 대한 인식실험의 결과를 나타낸다. 표에서 보는 바와 같이 모음 모듈의 평균 오인식률을 9.3%로서 우수한 성능을 보이는 반면, 무성음 모듈은 19.4%, 비음 모듈은 16.8%로서 비교적 높은 오인식률을 나타내고 있다. 음소 오인식은 단어의 시작 부분 및 끝 부분에서 추출한 음소에서 주로 많이 발생하였다. 이는 음성의 시작 부분 및 끝 부분에서 있을 수 있는 불완전한 포만트 구조에 기인한 것으로 생각된다.

표 2. 음소 인식실험의 결과(%)
Table 2. Experimental results(%) of phone recognition.

모듈	음소	오인식률	평균 오인식률	전체 오인식률
모음	아	6.5	9.3	15.2
	어	19.3		
	에	8.8		
	이	2.9		
	으	10.7		
	오	8.9		
	우	3.1		
	ㄹ	13.9		
비음	ㅁ	18.1	16.8	
	ㅇ	15.9		
	ㄴ	16.5		
무성음	ㄷ	19.6	19.4	
	ㅂ	18.8		
	ㅈ	20.8		
	ㅊ	19.1		
	ㅅ	18.1		
	ㅍ	20.3		
	ㄱ	19.7		
	ㅎ	18.9		

3. 단어인식 실험

제안한 인식 알고리듬의 성능을 평가하기 위해 단어음에 대한 인식실험을 수행하였다. 먼저, 각 음소별 최소 지속시간은 MLP의 출력값과 단어의 파형으로 부터 수 작업으로 측정하였다. 표 3에 보인 바와 같이 지속시간은 대략 1~12 프레임의 다양한 분포를 보인다.

단어인식 실험은 단어사전만을 이용한 알고리듬 1, 단어사전과 최소 지속시간을 함께 고려한 알고리듬 2, 그리고 기존의 연속 HMM(CHMM:Continuous Hidden Markov Model)[7]을 이용한 방법에 대하여 수행하였다. CHMM은 상태 수를 6으로한 모델로서, 이의 특징벡터(관측벡터)는 알고리듬 1 및 2의 것과 동일하다. 표 4(a)는 숫자음에 대한 인식결과로서, 알고리듬 1은 기존의 CHMM과 비슷한 성능을 보이고

있으나 알고리듬 2는 3.6%의 오인식률로서 알고리듬 1에 비해 오인식률이 1/3로 감소하였다.

표 3. 음소별 최소 지속시간
Table 3. Minimum duration of each phone.

모 음	아	어	에	이	으	오	우	ㄹ
	4	4	6	4	2	4	5	6
비 음	ㅁ	ㅇ	ㄴ					
	12	6	5					
무성음	ㄷ	ㅂ	ㅈ	ㅊ	ㅅ	ㅍ	ㄱ	ㅎ
	1	3	2	4	4	2	3	2

이 알고리듬에서는 어떤 음소의 지속시간이 최소값을 넘지 못하면 다른 음소로 천이할 수 없도록 제한한다. 그러므로 MLP의 분류 오류로 인해서 생길 수 있는 짧은 길이의 음소는 제거된다. 표 4(b)는 도시명에 대한 인식결과로서, 알고리듬 1과 2의 결과는 숫자음에 대한 인식결과에 비해서 매우 저조한 성능을 보이는 반면, 기존의 CHMM은 거의 비슷한 결과를 나타내었다. 알고리듬 1과 2는 음소단위로 단어를 인식하므로 음소의 수가 많은 단어에 대해서는 인식 성능이 다소 떨어지는 단점이 있다.

표 4. 인식 결과
(a) 숫자음의 오인식률 (b) 도시명의 오인식률
Table 4. Recognition results.
(a) Error rates of digits,
(b) Error rates of city names.

알고리듬1	10.3%
알고리듬2	3.6%
기존 CHMM	13.3%

(a)

알고리듬1	18.0%
알고리듬2	11.0%
기존 CHMM	12.2%

(b)

표 5는 숫자음과 도시명에 전체에 대한 대한 인식결과로서, 단어사전만을 이용한 알고리듬 1은 기존의 CHMM에 비해 저조한 성능을 보이지만, 최소 지속시간을 고려한 알고리듬 2에서는 7.3%로서 CHMM의 12.6%에 비해 5.3% 감소한 오인식률을 보였다. 그러므로 제안한 시스템은 기존의 CHMM에 비해 인식시의 계산량을 크게 줄이면서도 인식성능을 개선시킨다.

표 5. 숫자음과 도시명의 오인식률
Table 5. Error rates of digits and city names.

알고리듬 1	18.1%
알고리듬 2	7.3%
기존 CHMM	12.6%

그림 6은 MLP의 학습 오차에 따른 오인식률을 보인 것으로 대체로 학습 오차가 0.14 부근에서 오인식률이 최저가 됨을 알 수 있다. 학습 오차가 매우 큰 경우는 MLP가 음소인식을 제대로 수행하지 못하므로 오인식률이 증가되며, 학습 오차가 매우 작은 경우에는 MLP의 분류능력이 지나치게 예민해져서 음소 오인식이 발생할 경우 올바른 음소에 해당되는 출력값을 0으로 만들므로 인식 알고 리듬에서는 이를 충분히 고려할 수 없다. 그러므로 MLP 학습은 알고리듬이 대처할 수 있는 적절한 수준까지만 수행되어야 한다.

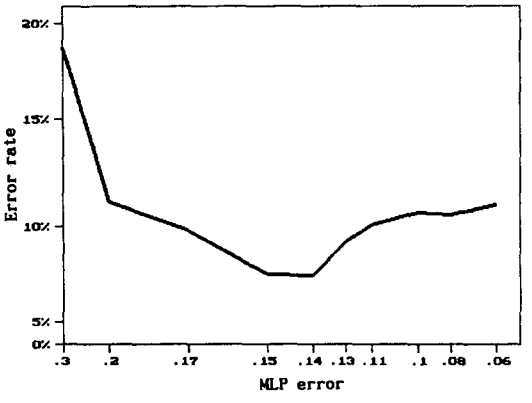

그림 6. MLP 오차에 따른 오인식률
Fig. 6. Recognition error for each MLP error.

그림 7은 단어에 포함된 음소 수에 따른 오인식률을 나타낸 것으로서, 음소의 수가 하나 또는 둘인 경우에는 거의 모두 인식되었다. 반면, 음소의 수가 3 이상일 경우에는 오인식률이 높아지고 있는데, 이는 주로 "전주"와 "청주" 같은 유사한 단어간에 오인식이 많이 발생하기 때문이다. 또 음소 수가 많은 단어에서는 화자에 따라 각 음소를 정확히 발음하지 않는 발성습관이나

음소간 조음효과 등의 원인으로 음소 고유의 특성이 사라지거나 특정한 음소는 누락되기도 하므로 오인식이 발생하는 것으로 생각된다.

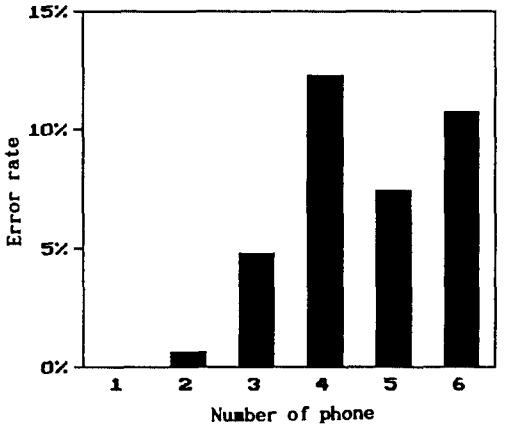

그림 7. 음소수에 따른 오인식률 분포
Fig. 7. Error distribution for each number of phone.

V. 결 론

본 논문에서는 고립단어를 음소단위로 나누어 인식하는 단어인식 시스템을 구성하였다. 음소단위의 음성인식은 조음효과나 화자의 발성습관에 따른 불분명한 음소 등의 영향으로 음소인식 자체가 어렵다. 그러므로 음소를 인식단위로 하는 단어인식은 매우 어려운 과제가 된다. 본 논문에서는 이러한 문제점을 해결하는 한 방안으로 단어사전을 이용한 네트워크 탐색 알고리듬과 음소의 지속시간 정보를 고려한 인식 알고리듬을 제안하였다. 이 알고리듬은 기준 모델이나 템플레이트를 이용하지 않고 단어사전에 의한 인식을 수행하므로 기존의 인식기에서 필요한 메모리 용량과 계산량을 크게 줄일 수 있다.

제안한 알고리듬의 성능을 확인하기 위해서 26개의 단어를 가지는 어휘에 대해 인식실험을 수행하였다. 그 결과, 숫자음의 오인식률은 3.6%, 도시명은 11.0%의 인식성능을 보였다. 그리고 전체 어휘에 대한 오인식률은 7.3% 로서 CHMM의 12.6% 에 비해 5.3% 개선된 인식성능을 보였다.

참 고 문 헌

[1] C. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithm for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 623-635, Dec. 1980.

[2] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Tech. J.*, vol. 64, pp. 1211-1234, July-Aug. 1985.

[3] B. H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic obsevations of Markov chain s." *AT&T Tech. J.*, vol. 64, pp. 1235-1249, July-Aug. 1985.

[4] R. L. Watrous, "Source decomposition of acoustic variability in a modular connectionist network," in *Proc. ICASSP*, pp. 129-131, 1991.

[5] 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," *B.S.T.J.*, vol. 62, pp. 1075-1105, Apr. 1983.

[6] P. L. Cerf, W. Ma, and D. V. Compernolle, "Multilayer perceptrons as lablers for hidden Markov models," *IEEE Trans. Speech Audio Proc.*, vol. 2, no. 1, part II, pp. 185-193, Jan. 1994.

[7] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1404-1413, Dec. 1985.

[8] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, pp. 4-22, Apr. 1987.

[9] A. Waibel, T. Hanazawa, G. Hinton, K.

Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 328-339, Mar. 1989.

[10] R. P. Lippmann, "Review of neural networks for speech recognition," *Readings in Speech Recognition*, Morgan Kufmann, Pub., 1990.

[11] P. D. Wasserman, *Neural Computing-Theory and Practice*, Van Nostrand Reinhold, New York, 1989.

[12] A. Waibel, H. Sawai, and K. Shikano, "Modularity and scaling in large phonemic neural networks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1888-1898, Dec. 1989.

[13] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297-315, Feb. 1975.

[14] 이한호, 정홍, "음절을 기반으로한 한국어 음성인식," 전자공학회논문지, 제 31권, B편, 제 1호, pp. 11-22, 1994. 1.

[15] J. Tebelskis and A. Wailbel, "Performance Through Consistency: MS-TDNN's for Large Vocabulry Continuous Speech Recognition", *Advanced in Neural Information Processing Systems V*, pp. 696-703, Dec. 1992.

저 자 소 개

李 基 熙(正會員) 第 32 券 B 編 第 5 號 參照
현재 대유공업전문대학 사무자동화과 전임강사

林 寅 七(正會員) 第 30 券 B 編 第 2 號 參照
현재 한양대학교 전자공학과 교수