퍼지패턴매칭에 의한 음성인식에 관한 연구

이 기 영 , 김 한 재 , 최 갑 석
명지 대학교 전자공학과

A Study on Speech Recognition Using Fuzzy Pattern Matching

Ki Young Lee , Han Jae Kim , Kap Seok Choi
Dept. Electronics Eng. Myong Ji Univ.

요 약

본 연구에서는 음성의 패턴작성법을 개선하고 음성인식율을 향상시키기 위하여 퍼지패턴매칭을 개선한 뉴럴-퍼지패턴매칭에 (a neural-fuzzy pattern matching) 의해 특정화자 고립단어인식을 수행하였다. 이 방법에서는 신경회로망의 연상기억에 의한 사상에 의해 패턴을 작성하여 주파수변동을 흡수하고 표준패턴과 선형매칭에 의해 유사도를 측정하여 인식하므로써 시간변동의 문제를 보완하였다. 또한, 이 방법에서 사용하는 특징파라미터는 2진화 스펙트럼이며, 유사도는 논리연산에 의해 측정되기 때문에 종래의 왜곡척도를 이용한 DTW 방법에 비해 기억용량과 계산량이 매우 적다.

이 방법의 인식성능을 평가하기 위하여 남여가 발성한 28 개의 도시명을 대상으로 인식실험을 수행한 결과, 신경회로망을 이용하지 않은 퍼지패턴매칭보다 고인식율을 얻을수있었으며,뉴럴-퍼지 패턴매칭에 의한 특정화자 고립단어인식의 우수성을 확인하였다.

1. 서 론

음성의 패턴인식에서 인식성능을 저하시키는 문제점으로는 음성의 시간변동과 주파수변동 등이 있다. 이러한 문제를 보완하는 방법으로 DTW 방법[2]과 멀티 템플릿[3]등이 개발되었으나, 모두가 대량의 기억용량과 계산량을 필요로 하는 단점이 있다.

최근, 인간 두뇌의 근사모델인 신경회로망과 애매성을 해결하려는 퍼지이론이 음성인식에 적용되면서 음성의 시간변동과 주파수변동 문제를 해결하기 위한 연구가 활발히 진행되고 있다. 신경회로망을 이용한 방법으로는 동적 프로그램과 결합한 DP, 시간지연에 따라 신경회로망을 적용한 TDNN 등이 있으며, 퍼지이론을 이용한 방법으로는 퍼지추론이나 퍼지패턴매칭에 의한 방법등이 있다.

그러나 이들 방법에서는 선형적인 대역통과필터분석, FFT 분석, 선형예측분석에 의해 추출한 특징을 입력으로 하여 인식을 수행하기 때문에 시간변동이나 주파수변동을 인식과정에서 흡수하지 않으면 안되므로, 많은 학습데이타를 필요로 하거나, 시행착오에 의한 인식방법의 개선 및 인식방법의 구조개선등이 필요하다. 이에 대처하기 위하여, 특징을 추출하는 전처리 과정에서 미리 시간변동이나 주파수변동 문제를 보완해 주려는 연구가 계속되고 있다.

본 연구에서는 음성인식을 수행하기 전에 신경회로망을 패턴작성 과정에 사용하여 주파수변동을 흡수하고, 인식과정에서 퍼지패턴매칭을 사용하여 시간변동의 문제를 보완하며 계산량과 기억용량이 적은 뉴럴-퍼지 패턴매칭 (a neural-fuzzy pattern matching) 에 의한 음성인식 방법을 제시하고자 한다.

본 연구의 뉴럴-퍼지 패턴매칭에 의한 음성인식 방법의 성능을 평가하기 위하여 28개의 도시명을 대상으로 음성인식을 수행하였으며, DTW 방법과 퍼지패턴매칭의 기억용량, 계산량 및 인식율과 비교, 검토하였다.

2. 음성의 특징파라미터 추출

음성이 입력되면 3.4 kHz 차단주파수의 저역통과필터를 통해 10 kHz 의 샘플링주파수로 A/D 변환하며, 특징파라미터로 구성된 패턴을 작성하기 위해 선형예측분석에 의해 스펙트럼을 추정한다. 스펙트럼의 시계열에 나타나는 음성의 피크주파수는 음원특성과 함께 성도의 전달특성을 수반하기 때문에 음성인식에서 중요한 특징파라미터의 역할을 한다.

따라서 본 연구의 음성패턴을 구성하는 특징파라미터는 스펙트럼시계열의 각 피크주파수를 청각특성에 근사한 1/3 옥타브 대역으로 변환하고 2 진화 처리한 2진화 스펙트럼이다. 표 1 에 1/3 옥타브 대역의 채널별 중심주파수를 보이고 있으며, 그림 1 에 선형예측분석에 의하여 특징파라미터를 추출하는 과정을 나타내었다.

그림 1 에서 음성신호가 입력되면 선형예측분석하여 스펙트럼을 추정하고 이 스펙트럼의 피크주파수를 피크피킹 알고리즘[9]에 의하여 검출한다. 한 스펙트럼의 피크주파수들을 청각특성에 근사시키기 위해 다음 식에 의해 피크주파수을 1/3 옥타브 대역[11]에 해당하는 채널로 변환한다.

$$k_j = INT\left[\, 3 \times \log_2 \frac{f_{pj}}{f_0} + 1 \right] \qquad (1)$$

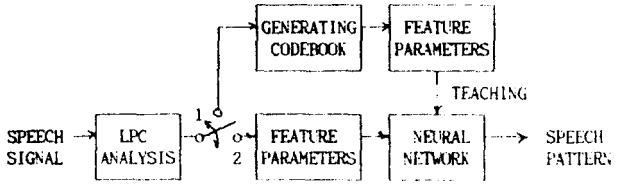
여기서 k_j 는 1/3 옥타브 대역의 18 개 대역 중에서 j 번째 피키주파수에 해당하는 대역의 채널을 나타내며 f_0 는 1 채널의 중심주파수 100 Hz 이며 f_{pj} 는 j 번째 피크주파수이고 INT [·] 정수화 함수이다.

그림 2 에 식(1)을 이용하여 음성의 스펙트럼을 2진화 스펙트럼으로 변환하는 예를 보이고 있다. 식(1)에 의해 피크주파수를 해당하는 1/3 옥타브 대역의 각 채널로 변환하고 그 채널을 1 로, 나머지 채널을 0 으로 표시하여 각 스펙트럼 단위로 특징파라미터를 추출하며 이와 같은 과정을 음성의 전 스펙트럼시계열에 수행한다.

3. 신경회로망에 의한 패턴작성

음성의 동일한 음운은 이상적으로 성도의 전달특성과 음원특성이 동일해야 하지만 같은 화자일지라도 성도의 순간적인 위치 변화와 유연성 때문에 동일한 음운의 스펙트럼이 변화하여 주파수변동[12]을 일으킨다. 본 연구의 패턴은 피크주파수의 위치에 의해 작성되기 때문에 주파수변동은 인식성능의 장애요인이 된다.

본 연구에서는 각 음운이 동일한 음원특성과 성도의 전달특성을 가지고있다면 같은 음운의 스펙트럼도 역시 동일해야 한다는 점에서 착안하여 신경회로망에 의해 유사한 스펙트럼들을 하나의 스펙트럼으로 사상해 주므로써 주파수 변동이 흡수된 패턴을 작성한다. 이에 따라 집단의 수는 음운의 종류수 이상으로 해 줄 필요가 있다. 유사한 스펙트럼들을 하나의 집단으로 보고 그 집단의 대표 스펙트럼을 생성하기 위해 집단화 알고리즘을 사용한다. 그림 3 은 신경회로망에 의해 패턴을 작성하는 구성도이다.



1 : TRAINING 2 : EXECUTING

그림 3. 신경회로망에 의한 패턴작성의 구성도
Fig. 3. Block diagram for making pattern using a neural network

그림 3 에서 신경회로망의 학습과정은 다음과 같다.
(a) 음성을 선형예측분석한 후, 집단화 알고리즘에 의해 음성의 코드북을 생성하며 각 집단의 대표 스펙트럼의 특징파라미터를 추출한다.
(b) 각 집단의 대표 스펙트럼의 특징파라미터를 교사응답으로 하고 그에 해당하는 집단 스펙트럼들의 특징파라미터들을 학습데이타로 하여 신경회로망을 훈련시킨다.

이상의 학습과정에 의해 신경회로망이 훈련되면 다음과 같은 수행과정에 의해 패턴을 작성한다.
(a) 입력된 음성으로 부터 선형예측분석에 의해 특징파라미터를 추출하고 신경회로망에 입력하면 해당하는 집단의 대표 특징파라미터를 출력한다.
(b) 위와 같은 과정을 음성의 전 시계열에 걸쳐 적용하여 각 집단의 특징파라미터로 구성된 패턴을 작성한다.

그림 4 에 신경회로망에 의해 패턴을 작성하는 예를 보이고 있다. 여기서 (a) 는 음성신호, (b) 는 음성신호의 스펙트럼 시계열, (c) 는 식(1)에 의한 2 진화 스펙트럼 패턴, (d) 는 신경회로망에 의해 작성한 7 진화 스펙트럼 패턴이다.

4. 뉴럴-퍼지 패턴매칭에 의한 음성인식

뉴럴-퍼지 패턴매칭에 의한 음성인식은 신경회로망 [연상기억 능력에 의해 패턴을 작성하여 주파수변동을 흡수하며, 퍼지패턴매칭[7]의 유사도측정을 이용하여 시간변동의 문제를 보완한다.

뉴럴-퍼지 패턴매칭에 의한 음성인식에서 표준패턴으로 등록된 단어들을 I = (i1, i2, ..., iM), 단어명 ik 를 발성했을 때 특징파라미터로 구성된 패턴을 X = (x1, x2, ..., xN), 각 단어패턴의 귀속성을 나타내는 멤버쉽함수를 Y = (m1, m2, ..., mM) 이라 하자. 여기서 멤버쉽함수 mj 는 단어패턴 xj 의 귀속등급을 나타내므로 각 단어 마다 여러 번 발성하여 작성한 패턴들을 종합하여 작성한다. 그러나 각 패턴의 주파수축의 길이는 일정하나 시간축의 길이는 서로 다르기 때문에 패턴을 종합하기 전에 선형신축 방법[10]에 의해 길이를 정규화시킨다.

그림 5 은 신경회로망을 전처리 과정에서 사용하여 주파수변동을 흡수하며 표준패턴을 멤버쉽함수로 작성하여 시간변동 문제를 보완하는 뉴럴-퍼지 패턴매칭과 사대유사도결정규칙에 의해 음성을 인식하는 구성도이다. 그림 5 에서 신경회로망은 3-1절의 학습과정에 의해 훈련된 것이며, 음성이 입력되면 신경회로망을 거쳐 특징파라미터로 구성된 음성의 패턴을 작성한다. 또한, 같은 음성의 패턴들을 종합한 멤버쉽함수로 표준패턴을 작성하며 시험패턴과 표준패턴 사이의 유사성을 측정하여 최대유사도결정규칙에 의해 음성을 인식한다. 시험패턴을 y (y ∈ X), j 번째 등록된 단어의 표준패턴인 멤버쉽함수를 aj 라 할때 이들 사이의 유사도[7]를 측정하는 식은 다음과 같다

$$S_{jy} = \frac{P_{jo}}{P_{jo} + P_{jo}} \cdot \frac{P'_{jo}}{P_y + P'_{jo}} \qquad (2)$$

여기서, $P_{jo} = \sum_t \sum_f m_j(f,t)$ (3)

$P_y = \sum_t \sum_f y(f,t)$

$P_{jo} = \sum_t \sum_f m_j(f,t) \cap y(f,t)$

$P'_{jo} = \sum_t \sum_f m_j(f,t) \circ y(f,t)$

또한 시험패턴 y 와 멤버쉽함수 m_j 는 주파수 f 와 시간 t 의 2 차원 패턴이며 ∩ 는 두 패턴 사이의 논리곱이고 ∘ 는 y 와 m_j 의 α-cut 사이의 논리곱이다. 즉,

$$m_j(f,t) \cap y(f,t) = \begin{cases} 1, & m_j(f,t) \geq 1 \text{ and } y(f,t) = 1 \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

$$m_j(f,t) \circ y(f,t) = \begin{cases} 1, & m_j(f,t) \geq \alpha \text{ and } y(f,t) = 1 \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

여기서 α 는 1 부터 멤버쉽함수의 최대치 사이에서 선택한다.

다음 식은 결정규칙에서 최대유사도를 갖는 표준패턴의 단어명을 선택하여 미지단어를 인식하는 식이다.

$$j^* = \arg \max_{1 \leq j \leq n} [S_{jy}] \qquad (6)$$

여기서, n 은 등록 단어의 수이며, S_{jy} 는 유사도 측정식이다.

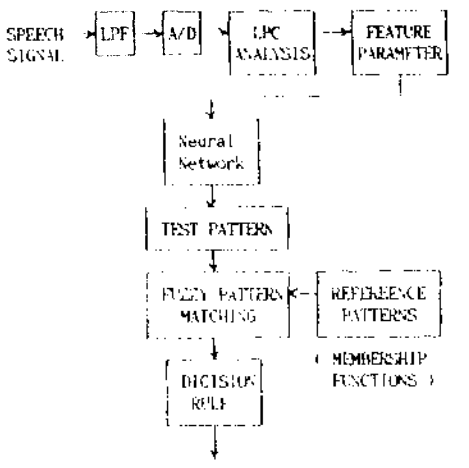


그림 5. 뉴럴-퍼지 패턴매칭에 의한 음성인식 구성도
Fig. 5. Block diagram of speech recognition using Neural-Fuzzy pattern matching

5. 실험 및 고찰

5-1. 음성 데이타

음성 데이타는 DDD 전화 도시명 28 개 (직할시 이상 6개 + 경기도내 22 개 = 28 개) 를 방음장치가 되지 않은 실험실에서 20 대 남성 2 인과 여성 1 인이 5 회씩 발성 (28개 × 3인 × 5회 = 420개) 하여 실험데이타로 사용하였으며, 표 2 와 같은 조건에 의해 음성을 분석하였다.

뉴럴-퍼지 패턴매칭에서 신경회로망을 훈련시키기 위해 각 화자가 3 회씩 발성한 DDD 전화 도시명 (28개 × 3인 × 3회 = 252개) 을 선형예측분석하여 특징파라미터를 추출하고 이들을 학습데이타로 하였으며, LBG 집단화 알고리즘에 의해 코드북을 생성하였고 이 코드북의 각 스펙트럼으로 부터 특징파라미터를 추출하여 교사응답으로 사용하였다. 여기서 신경회로망은 역전파 학습알고리즘에 의해 학습되었으며, 입력층과 출력층의 유니트 수가 각각 18개, 은둔층은 5개로 구성하였다.

표 1. 1/3 옥타브 대역의 채널별 중심주파수[11]
Table 1. Center frequency of a channel of 1/3 octave bands

CHANNEL	CENTER FREQUENCY [Hz]
1	100
2	125
3	160
4	200
5	250
6	317
7	400
8	504
9	635
10	800
11	1,008
12	1,270
13	1,600
14	2,016
15	2,540
16	3,200
17	4,032
18	5,000

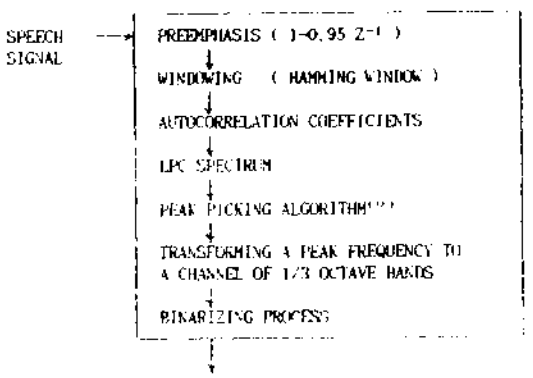


그림 1. 선형예측분석에 의한 특징파라미터의 추출과정
Fig. 1. Procedure for extracting feature parameters using linear prediction analysis

표 2. 분석 조건
Table 2. Analysis conditions

Cut-off frequency	3.4 kHz
Sampling frequency	10 kHz
A/D resolution	12 bits
Window length	30 ms
LPC orders	14

3-2. 인식율의 비교

뉴럴-퍼지 패턴매칭에 의한 인식성능을 검토하기 위하여, 신경회로망을 이용하지 않은 퍼지패턴매칭에 의한 인식율과 비교, 검토하였다.

(1) 퍼지패턴매칭에 의한 인식실험

신경회로망을 이용하지 않고 패턴을 작성하는 첫번째 방법은 먼저 선형예측에 의해 추정한 스펙트럼으로 부터 피크주파수와 대역폭을 모두 검출하고 1/3 옥타브 대역의 채널로 변환하여 피크주파수와 대역폭에 해당하는 채널을 1 로, 나머지 채널을 0 으로 표시하므로써 패턴을 작성한다. 멤버쉽함수는 이 방법으로 작성된 패턴의 3 회분을 선형중첩하여 작성하며 표준패턴으로 사용한다. 또한, 식(2)와 식(6)에 의해 음성인식을 수행한다. 표 3 에는 피크주파수와 대역폭으로 패턴을 작성하는 첫번째 경우 α 를 변수로 하여 인식율을 표시하였다.

표 3. 첫번째 방법에 의한 인식율 [%]
Table 3. Recognition rates by 1st. method [%]

SPEAKER \ α	1	2	3
MA	75.0	91.1	83.9
MB	85.7	92.9	73.2
FC	58.9	92.9	82.1
AVG.	73.2	92.3	79.7

표 3 의 인식율에서 α 값이 2 일때 평균 92.3%의 인식율을 얻었으며 오인식은 조음위치가 동일한 모음을 공통적으로 포함한 도시명 「부산,운산」, 「안산,안성」, 「인천,인천」 사이에서 주로 발생하였다. 이러한 오인식이 발생한 것은 1/3 옥타브 대역이 저대역 주파수에서는 좁고 고대역 주파수로 갈 수록 넓어 지기 때문에 저대역 주파수 특성이 비슷한 모음을 포함한 음성은 피크주파수와 그 대역폭으로 표현된 패턴이 서로 유사하여 오인식이 발생하였다.

신경회로망을 이용하지 않는 두번째 방법에서는 첫번째 방법의 오인식을 피하기 위하여 피크주파수에 해당하는 채널만을 1로 표시하고 나머지 채널을 0으로 표시하여 패턴을 작성하고 식(6)에 의해 음성인식을 수행하였다. 표 4 에는 피크주파수만으로 패턴을 작성한 경우 α 값을 변수로 하여 인식율을 표시하였다.

표 4 의 인식율에서 α 값이 2 일때 첫번째 방법에서 발생했던 오인식을 줄일 수 있었으며, 평균 94.1%의 인식율을 얻으므로써 신경회로망을 이용하지 않은 첫번째 방법의 경우보다 인식성능을 향상시킬 수 있었다.

표 4. 두번째 방법에 의한 인식율 [%]
Table 4. Recognition rates by 2nd. method [%]

SPEAKER \ α	1	2	3
MA	94.6	96.4	71.4
MB	89.3	92.9	62.5
FC	91.1	92.9	73.2
AVG.	92.5	94.1	69.0

(2) 뉴럴-퍼지 패턴매칭에 의한 인식실험

신경회로망을 이용하지 않은 인식방법에서 피크주파수만을 패턴으로 작성하는 두번째 방법에서 인식성능을 향상시킬 수 있었으므로, 본 절의 뉴럴-퍼지 패턴매칭에서는 피크주파수만으로 표현된 특징파라미터의 시계열을 작성하며 신경회로망에 의해 전처리를 행하여 패턴을 작성하여 음성인식을 수행한다.

이 방법에서 신경회로망에 의해 사상되는 집단의 수는 음운의 수와 일치시켜야 하지만 본 연구에서는 유사한 스펙트럼들을 한 집단으로 가정하므로 집단의 수를 변화시키면서 인식율을 비교하였다.

표 5 에는 각각 집단의 수가 16, 32, 64 일 때, α 값을 변수로 한 인식율을 나타내었다. 여기에서 코드북사이즈가 32 이고 α 값이 2 일 때 모음이 유사한 도시명의 오인식이 감소되었을 뿐만 아니라, 이때 남성화자 MA

MB 는 각각 98.2%와 96.4%, 여성화자 FC 는 94.6% 로 평균 96.4%의 인식율을 나타내므로써 퍼지패턴매칭에 의한 방법보다 오인식을 약 1/2 배 감소시킬 수 있었다.

그런데, 집단의 수가 16인 경우에는 α 가 2 일때 92.9% 의 평균 인식율을 얻으므로써 집단의 수가 32인 경우보다 인식율이 낮았다. 이것은 실험에 사용할 도시명이 자음 14개와 기본모음 8개로 구성되어 있으므로 신경회로망에 의해 사상되어야 할 집단의 수는 적어도 22개 이상이 필요한데 대해 그수가 모자르기 때문에 서로 달라야 할 음운이 서로 같은 음운의 스펙트럼으로 사상되었기 때문이라 생각된다.

또한 집단의 수가 64 인 경우, α 가 2일 때 평균 95.8%의 인식율을 얻었다. 여기서 집단의 수가 32 일 때보다 인식율이 떨어지는 것은 집단의 수가 음운의 수보다 너무 많기 때문에 주파수 변동분까지 집단으로 포함되어 신경회로망에 의한 사상으로 주파수 변동을 흡수하지 못한 것으로 생각된다.

집단의 수가 32 인경우 α 가 2일 때 가장 높은 인식율을 얻을 수 있었던 것은 집단의 수가 음운의 수와 가장 가깝기 때문에 같은 음운을 하나의 스펙트럼으로 나타낼 수 있음으로 해서 주파수 변동을 흡수했기 때문이라 생각된다.

6. 결 론

본 연구에서는 음성의 주파수변동과 시간변동 문제를 보완하기 위하여 신경회로망과 퍼지이론을 결합한 뉴럴-퍼지 패턴매칭에 의한 특정화자 고립단어인식을 수행하였다. 이 방법을 평가하기 위하여 28개의 도시명을 대상으로 음성인식 실험을 수행한 결과 다음과 같은 결론을 얻었다.

(1) 뉴럴-퍼지 패턴매칭에 의한 음성인식율은 96.4% 였으며, 퍼지패턴매칭보다 오인식을 감소시켰다.

(2) 뉴럴-퍼지 패턴매칭에서 종래의 음성인식방법보다 작은 기억용량과 계산량으로 DTW 방법과 유사한 인식율을 얻으므로써 이 방법의 우수성을 확인할 수 있었다.

참고 문헌

[1] W.A.Lea,Tends in speech recognition , Prentice-Hall, Inc. 1980

[2] H.Sakoe,S.Chiba,"Dynamic programming algorithm optimization for spoken word recognition,"IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-26, Feb.1978

[3]S.E.Levinson,L.R.Rabiner,A.E.Rosenberg,J.G.Wilpson,"Interactive clustering techniques for selecting independent reference templates for isolated word recognition,"IEEE Trans. Acoust. Speech, Signal Processing, Vol.ASSP-27,Apr. 1979

[4] H.Sakoe,R.Isotani,K.Yoshida,"Speaker-Independent word recognition using dynamic programming neural network,"ICASSP 89, S1.8,1989

[5] A.Waibel,T.Hanazawa,G.Hinton,K.Shikano,K.Lang," Phoneme recognition using Time-Delay Neural Networks ," IEEE Trans.Acoust.Speech,Signal Processing, Vol.ASSP-37, 1989

[6] 崔元奎,秋月影雄,"ファシィ推論による母音認識と韓國語連續音聲への應用," 電學論 C, 111-5, 1991

[7] J.Fuzimoto, T.Nakatani, M.Yoneyama, " Speaker-independent word recognition using fuzzy pattern matching, " Fuzzy Sets and Systems 32, North-Holland, 1989

[8] S.Nakanishi,T.Takagi,"Pattern recognition by neural network and fuzzy inference,"Proc. ICFAN, 1990

[9] 崔元奎,李義 ,秋月影雄,"韓國語音聲認識のための改良形ホルマントトラッキングと數子音聲認識への應用,"電學論 C, Vol. 109-C, No.10, 1988

[10] 松本博,中川正雄,来山正秀,"ローカルピーク荷重平均辭書を用いた不特定話者單語音聲認識,"電子通信學會論文誌,Vol.J68-A, No.1, 1985

[11] R.A.Cole, Perceptron and production of fluent speech , Laurence Erlbaum Associates, Inc.,1980

[12] W. J. Hess," Algorithms and devices for pitch determination of speech signal," J.P.Haton (ed.) ,Automatic Speech Analysis and Recognition, 1982

[13] A.H.Gray,Jr.,J.D.Markel,"Distance measure for speech processing,"IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-24,Oct.1976

[14] 이기영,최갑석,"사상코드북을 이용한 화자적응 한국어 숫자음 인식에 관한 연구,"한국음향학회지 제9권 5호 1990

표 5. 뉴럴-퍼지 패턴매칭에 의한 인식율
Table 5. Recognition rates by Neural-Fuzzy pattern matching

[%]

α / CLUSTERS	1			2			3		
SPEAKER	16	32	64	16	32	64	16	32	64
MA	92.9	92.9	92.9	94.6	98.2	98.2	73.2	78.6	73.2
MB	91.1	92.9	92.9	91.1	96.4	96.4	89.3	82.1	89.3
FC	91.1	92.9	91.1	92.9	94.6	92.9	82.1	75.0	78.6
AVG.	91.7	92.9	92.3	92.9	96.4	95.8	81.5	79.8	79.7