論文 89-26-7-18

화자인식을 위한 음성 요소들의 성능분석 및 새로운 판단 논리

(Performance Analysis of Speech Parameters and a New Decision Logic for Speaker Recognition)

李 赫 宰*, 李 秉 基*

(Hyuk Jae Lee and Byeong Gi Lee)

要 約

본 논문에서는 화자인식 시스템의 인식율 향상을 도모하기 위하여 요소의 선택 및 판단 논리의 문제를 고찰하였다. 또한 화자인식 실험을 수행하는 과정에서 기준패턴의 작성이 인식율에 어떠한 영향을 미치는 가를 아울러 검토해 보았다. LPC, PARCOR 계수, LPC-cepstrum 계수등을 인식 요소로 사용하여 화자확인 오차율을 측정한 결과, 기준 패턴의 작성방법에 관계 없이 LPC-cepstrum계수의 성능이 LPC나 PARCOR 계수의 성능에 비해 우수한 것으로 나타났다. 또 화자인식율을 향상시키기 위하여 일반화된 거리 개념을 도입한 새로운 판단 논리를 제안하였다. 제안된 판단 논리는 기준화자 및 외부화자의 통계적 성질을 동시에 고려하여 각 요소들에 서로 다른 가중치를 둔다는 점이 기존의 방법들에 비해 다르다. 화자확인 실험결과 제안된 판단 논리를 적용한 경우가 기존의 방법들에 비해서 인식율이 향상된 것을 관찰할 수 있었다.

Abstract

This paper discusses how to choose speech parameters and decision logics to improve the performance of speaker recognition systems. It also considers the influence of the reference patterns on the speaker recognition. It is observed from the performance analysis based on LPCs, PARCOR coefficients and LPC-cepstrum coefficients that LPC-cepstrum coefficients are superior to the others in speaker recognition without regard to the reference patterns. In order to improve the recognition performance, a new decision logic is proposed based on a generalized-distance concept. It differs from the existing methods in that it considers the statistics of customers and impostors at the same time. It turns out from a speaker verification test that the proposed decision logic ferforms better than the existing ones.

*正會員, 서울大學校 電子工學科
(Dept. of Elec. Eng., Seoul Nat'l Univ.)
接受日字 : 1989年 2月 24日

I. 서 론

화자인식(speaker recognition)은 음성인식의 한분야로서 음성에서 추출한 정보로 부터 화자의 신원을 인

식하는 것을 주목적으로 삼는다. 현재까지 연구되어온 화자 인식 기법은 크게 두가지로 나눌 수가 있다. 첫째는 패턴정합(pattern matching) 방법이다.[1] 이것은 각 화자를 대표할 수 있는 패턴들을 미리 작성한 다음, 시험패턴과 기준패턴 사이의 유사도를 측정하여 시험패턴의 신원을 파악하는 방법이다. 둘째는 통계적 성질을 이용한 분류 방법으로서 각 화자에서 추출한 음성요소(speech parameter)들을 오랜 시간동안 관찰하여 통계량을 구한 후 이것으로 신원을 파악하는 방법이다.[2] 패턴정합 방법은 통계적 방법에 비해 계산량은 많지만 높은 인식율을 얻을 수 있다는 장점 때문에 많은 연구가 수행되어 왔다.[1,5~7]

패턴정합을 이용한 화자인식 시스템에서 인식율을 향상시키기 위해서는 특히 다음 사항들을 고려하여야 한다. 먼저, 어떠한 음성요소를 선택하여 화자인식에 사용할 것인가 하는 요소선택의 문제가 고려되어야 한다. 즉, 화자인식에 유용한 음성요소들을 추출하고 이들의 성능을 분석 검토하는 연구들이 우선 필요한 것이다.[3~5] 또한, 시험용 데이타와 기준패턴 사이의 유사도를 측정하여 이것으로 부터 화자의 신원을 결정하는 문제가 고려되어야 한다. 이때, 각 요소들의 통계적 성질을 고려하기 위해서는 적절한 가중치를 둔 거리를 사용할 필요가 있다.[1,6,7]

본 논문의 목적은 윗 두가지 관점에서 부터 화자인식을 도모하는 방안을 모색하는데 있다. 먼저, 화자인식 요소로써 선형예측 계수(linear prediction coefficent : LPC), PARCOR(partial correlation) 계수, 선형예측 계수를 변환한 cepstrum(LPC-cepstrum) 계수 등을 택하고, 이에 각종 기준패턴 작성 방법을 적용하여 이들의 인식 성능을 비교 검토함으로서, 어떠한 요소 및 어떠한 기준패턴이 화자인식에 적합한 가를 고찰해 보도록 하겠다. 이어서 음성요소를 결합하여 인식 판단을 내리기 위한 새로운 판단 논리를제안하고, 인식실험을 통해서 이 판단논리의 성능을 분석해 보도록 하겠다. 본 논문에서 LPC, PARCOR 계수 및 LPC-cepstrum 계수를 사용한 것은 이들이 새로운 요소라거나 인식 성능이 좋은 것들이기 때문이 아니라, 이들을 토대로 각종 패턴작성 방법에 따른 이들 성능을 분석 검토하고 또, 뒤이어 제안한 새로운 판단 논리를 시험하기 위한 측정 수단으로 삼기 위한 것이다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서 본 연구에서 사용될 화자인식 시스템에 관해서 설명하고, 제Ⅲ장에서는 화자인식을 위한 음성요소들 및 각종 패턴들의 인식성능을 분석하도록 하겠다. 이를 토대로 제Ⅳ장에서는 화자인식을 위한 새로운 판단논리를 제안하고, 인식실험을 통하여 제안된 방법의 성능을 분석 검토해 보도록 하겠다.

Ⅱ. 화자인식 시스템

본 연구에 사용한 화자인식 시스템의 전반적인 구성은 그림 1과 같다. 먼저 전처리 과정을 통해 입력 신호들 가운데 불필요한 부분을 제거하고, 화자인식을 위한 요소들을 추출한 후, 이 요소들을 시간에 따라 나열하여 패턴을 작성한다. 이렇게 작성된 패턴은 이어서 시간 보정을 거치게 된다. 학습과정에서는 시간 보정이 된 여러 패턴을 조합하여 기준 패턴을 작성하고, 시험과정에서는 미리 작성된 기준패턴과의 유사도를 측정한다. 끝으로, 측정된 유사도에 판단논리를 적용시켜서 화자의 신원을 판단한다. 본 논문에서 사용한 바의 각 처리과정은 다음과 같다.

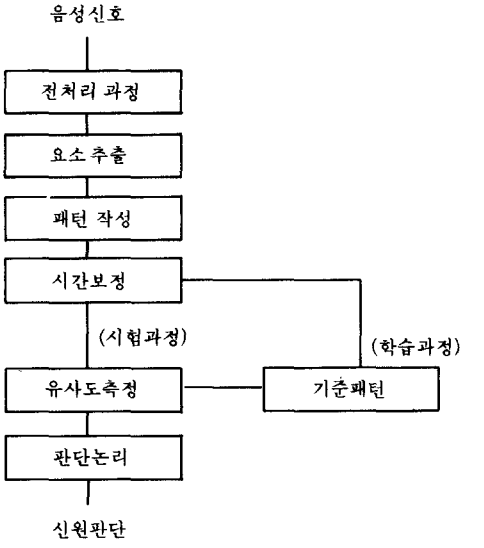


그림 1. 화자인식 시스템의 구성
Fig. 1. Organization of speaker recognition system.

1. 전처리 과정

입력된 아날로그 음성신호를 먼저 4KHz 저역여파(lowpass filtering) 시킨 후, 10KHz로 표본화(sampling)하였다. 이어서 직류성분을 제거하고, 에너지를 정규화(normalize)한 후 에너지와 영교차율을 측정하고, 이들을 이용하여 음성신호의 묵음 구간을 제거

하였다. 이러한 전처리과정을 통하여 묵음 구간이 제거된 신호로 부터 화자인식을 위한 요소들을 추출하였다.

2. 요소의 추출 및 패턴의 작성

화자인식을 위한 음성요소들은 묵음구간이 제거된 상태에서 각 프레임별로 추출하였다. 이때 음성요소로서 LPC, PARCOR 계수 및 LPC-cepstrum 계수 등을 사용하였다. 각 프레임의 길이는 그림 2에서 보는 것과 같이 25.6ms로 하고, 12.8ms씩 중첩(overlap)시켰다.

한 프레임에서 총 M개의 요소를 추출하여 이것으로 하나의 벡터를 구성하였다. 이 때 i번째 프레임에 추출한 M개의 요소를 $p_{i1}, p_{i2}, \cdots, p_{iM}$라고 하면, 이것으로 i번째 요소벡터 $\vec{p}_i$를 구성하였다. 즉,

$$\vec{p}_i = (p_{i1}, p_{i2}, \cdots, p_{iM})^t. \quad (1)$$

하나의 음성 데이타가 총 N개의 프레임으로 구성된다면, 이것을 나열하여 요소행렬 P를 구성할 수가 있다. 즉,

$$P = [\vec{p}_1, \vec{p}_2, \cdots \vec{p}_N] = \begin{bmatrix} p_{11} & p_{21} \cdots p_{N1} \\ p_{12} & p_{22} \cdots p_{N2} \\ - & - \quad - \\ - & - \quad - \\ p_{1M} & p_{2M} \cdots p_{NM} \end{bmatrix}$$

이렇게 작성된 M×N 요소행렬 P를 본 논문에서는 패턴이라고 부르기로 한다. 본 연구에서는 M=12로 두었으며, N은 55에서 95사이의 값이 되었다.

3. 시간보정 및 기준패턴의 작성

각 패턴간의 시간보정을 위해서는 DTW (dynamic time warping) 방법을 사용하였다.[9] 다음으로 시간

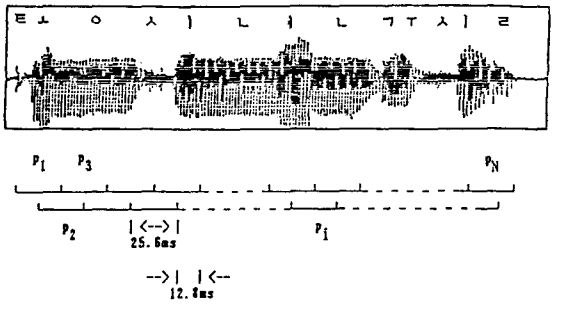


그림 2. 음성프레임과 요소벡터
Fig. 2. Speech frames and parameter vectors.

보정이 된 패턴을 사용하여 기준패턴을 작성하였다. 이를 위해서는 먼저 기준 화자의 데이타를 L번 받아들이고, 그 각각으로 부터 요소패턴을 하나씩 작성한다. 즉, 식(2)에 설명한 P와 같은 행렬을 만든것이다. 본 논문에서는 L=1, 5, 25로 하였다. 이러한 L개의 행렬 P로 부터 기준패턴을 작성하기 위해서는, 우선 각 패턴사이의 시간보정을 한 다음 전체의 산술 평균을 취하였다. 즉, 시간보정을 한 k번째 요소패턴을 P(k)라고 하면, 기준패턴 R은 L개의 P(k)의 산술평균으로서, R의 요소 r_{ij}와 P(k)의 요소 $p_{ij}(k)$간에는

$$r_{ij} = \frac{1}{L}\sum_{k=1}^{L} p_{ij}(k) \quad (3)$$

의 관계가 성립한다.

4. 유사도의 측정

두 패턴사이의 유사도를 측정하기 위하여, 서로 대응되는 프레임에서 추출한 요소들끼리 비교하여 거리를 구한 후, 이것을 전체 프레임에 대해서 평균하였다.

이를 수식으로 표현하면 다음과 같다. 기준패턴과 시험패턴을 각각 R, T라고 하고, 이들의 i번째 프레임의 요소벡터를 각각 $\vec{r}_i, \vec{t}_i$라고 하면, 즉,

$$R = [\vec{r}_1, \vec{r}_2, \cdots, \vec{r}_N], \quad T = [\vec{t}_1, \vec{t}_2, \cdots, \vec{t}_N], \quad (4)$$

이면, 이중 i번째 프레임간의 유클리드 거리 $d(t_i, r_i)$는

$$d(\vec{t}_i, \vec{r}_i) = (\vec{t}_i - \vec{r}_i)^t (\vec{t}_i - \vec{r}_i) = \sum_{k=1}^{M} (t_{ik} - r_{ik})^2 \quad (5)$$

이다. 모든 프레임에 대하여 이 거리의 평균을 취한 것을 두 패턴사이의 거리 D(T, R)로 정의하면,

$$D(T, R) = \frac{1}{N}\sum_{i=1}^{N} d(\vec{t}_i, \vec{r}_i) \quad (6)$$

이 된다. 그러므로, 식(5)와 식(6)을 결합하면, 결국

$$D(T, R) = \frac{1}{N} tr[(T-R)^t (T-R)] \quad (7)$$

의 관계를 얻는다. 이렇게 구한 거리 D(T, R)이 작으면 유사도는 크고, 거리가 크면 유사도는 작게 된다.

5. 판단 논리

화자인식의 최종 단계에서는 기준패턴과 시험패턴 사이의 거리를 학습과정을 통해서 정해진 문턱값

(threshold)과 비교하므로서 화자를 판단하였다. 즉, 문턱값을 D_{th}라 라고 하면,

$D(T,R) \leq D_{th}$이면 기준화자
$D(T,R) > D_{th}$이면 외부화자　　(8)

의 관계로서 기준화자 여부를 판단한 것이다. 이 때 나타나는 오판확률을 본 논문에서는 화자확인 실패율이라고 부른다.

Ⅲ. 화자인식을 위한 요소들의 성능 분석

화자인식을 위해서는 기본주파수, 포만트주파수, LPC, PARCOR 계수, LPC-cepstrum 계수, 에너지, 영교차율 등이 주로 사용된다. 이들중 LPC와 기본주파수가 화자인식에 유용하고, 특히 LPC를 변환한 LPC-cepstrum 계수가 우수한 것으로 알려져 있다.[5,6] 본 논문에서는 LPC, PARCOR 계수 및 LPC-cepstrum 계수를 택하여 화자인식 성능을 분석하고, 이 때 기준패턴 작성방법에 따른 화자확인 오차율을 검하여 검토하도록 하겠다.

1. 기준패턴의 작성

같은 화자가 동일한 발음을 하더라도 상황에 따라 패턴이 조금씩 달라지므로 다양한 상황을 고려하기 위해서는 많은 데이타를 바탕으로 기준패턴을 작성하여야 하겠다. 그러나 데이타가 너무 많은 경우에는 smoothing 작용 때문에 기준패턴의 특징이 상실될 수도 있다. 그러므로, 상황에 따른 변화와 특징의 유지를 모두 고려해 주기 위해서는 기준패턴을 작성하기 위한 데이타의 양이 조절될 필요가 있게 된다.

본 논문에서는 기준화자 5명의 각각에 대하여 25개의 데이타를 취해서 이들로 부터 4가지 종류의 기준패턴을 작성하였다. 이 25개의 데이타는 1주일에 5개씩 5주동안 받아들인 것이다. 이 때, 기준패턴은 12차의 LPC, 12차의 PARCOR 계수 및 12차의 LPC-cepstrum 계수들 각각에 대해서 별개로 정해지며 그 작성방법은 다음과 같다.

기준패턴 1은 25개의 데이타 각각을 기준패턴으로 삼아 인식 성능분석을 했을 때 가장 좋은 결과를 나타낸 것을 한개 택한 것이다.

기준패턴 2는 25개의 데이타 각각을 기준패턴으로 삼아 인식성능을 분석하고, 그 결과를 각 주별로 평균을 취했을 때 가장 좋은 성능을 나타낸 주의 평균 데이타이다.

기준패턴 3은 각 주에서 가장 좋은 성능을 보인 데이타를 1개씩을 택해서 이들의 평균을 취한 것이다.

기준패턴 4는 인식성능 실험없이 25개 데이타 모두의 평균을 취한 것이다.

이들 각 경우에 있어서 평균은 식(3)과 같이 취하였으며, 또한 시간보정 병행처리 되었다. 이 때 시간처리를 위해서는 DTW를 사용하였다.

2. 음성요소들의 성능분석 실험

본 연구에서는 "통신연구실"이라는 발음을 사용하여 5명의 기준화자를 두고, 그 각각에 대하여 화자확인 실험을 수행하였다. 이 때 기준화자 5명 각각에 대해서 기준화자 자신의 데이타 25개와 외부화자의 데이타 25개를 취해서 화자확인 시험용 데이타로 삼았다. 이 때 기준화자의 데이타 25개는 학습시 기준패턴을 작성하기 위해서 사용했던 데이타들과는 별개의 것들로 하였다.

시험용 데이타 50개의 각각에 대해서, 12차의 LPC 12차의 PARCOR 계수 및 12차의 LPC-cepstrum 계수들을 추출하고, 이들을 각각 4가지 종류의 기준패턴들과 비교하여 식(7)과 같이 거리를 측정한 후, 문턱값의 변화에 따라 식(8)과 같은 논리로 FR/FA 곡선을 그렸다. 이 때 FR(false reject)은 기준화자의 데이타인데도 불구하고 외부화자라고 오판할 확률을 나타내고, FA(false accept)는 외부화자의 데이타인데도 기준화자라고 오판할 확률을 나타낸다. FR/FA 곡선은 이러한 두가지 확률곡선을 동일 좌표위에 도시한 것으로서, 두 곡선이 교차하는 지점의 오차율이 화자확인 오차율이 된다.

3. 음성요소들의 성능분석 결과

표 1은 12차의 LPC를 음성요소로 사용하고 4가지 기준패턴을 작성하여 인식실험을 한 결과이다. 표 2와 표 3은 12차의 PARCOR 계수 및 LPC-cepstrum 계수를 사용하여 마찬가지로 인식실험을 수행한 결과이다. 이 표로 부터 우선 패턴에 무관하게 LPC-cepstrum 계수가 성능이 가장 좋고, 다음으로 PARCOR 계수가 좋으며, LPC는 상대적으로 나쁜 것을 살펴볼 수 있다. 이는 영문을 바탕으로 실험했던 기존의 연구결과들(참고문헌[11])이 한국어에 있어서도 부합됨을 보이는 것이다.

또한 화자에 따른 인식율의 차이가 큰 것을 관찰할 수 있다. 예를들어 화자 2, 3 및 4의 경우에 있어서는 LPC-cepstrum 계수를 사용하면 거의 완전한 화자인식이 가능하지만, 화자 1, 5의 경우에 있어서는 오차율이 크다. 이는 각 화자의 발음 습관이나 그 특이성이 화자확인 인식율에 큰 영향을 미친다는 사실을 보여주는 예이다.

또한 기준패턴에 따라서 성능이 차이가 많이 나는 화자가 있음도 관찰할 수 있다. 예를들어 표 4에서

표 1. LPC 테스트의 화자확인 오차율
Table 1. Speaker verification error rate of the LPC test.

화자	패턴 1	패턴 2	패턴 3	패턴 4	평균
화자 1	24	23	27	25	25
화자 2	3	8	6	7	6
화자 3	12	6	11	6	9
화자 4	2	6	2	18	7
화자 5	25	22	29	19	24
평균	13	13	15	15	14

표 2. PARCOR 테스트의 화자확인 오차율
Table 2. Speaker verification error rate of the PARCOR test.

화자	패턴 1	패턴 2	패턴 3	패턴 4	평균
화자 1	6	10	14	6	9
화자 2	0	0	0	0	0
화자 3	13	6	4	5	7
화자 4	3	0	5	0	2
화자 5	20	16	18	10	16
평균	8	6	8	4	7

표 3. LPC-cepstrum 테스트의 화자확인 오차율
Table 3. Speaker verification error rate of the LPC-cepstrum test.

화자	패턴 1	패턴 2	패턴 3	패턴 4	평균
화자 1	6	6	2	10	6
화자 2	0	0	0	0	0
화자 3	0	1	0	0	0
화자 4	0	0	2	0	1
화자 5	13	10	16	10	12
평균	4	3	4	4	4

화자1과 5는 기준패턴에 따라 인식율의 차이가 많이 난다. 또, 가장 우수한 성능을 보이는 기준패턴이 요소에 따라 달라지는 것을 관찰할 수 있다. 극단적인 예로 화자 4의 경우 LPC를 위해서는 기준패턴 1과 3이 우수한 성능을 보이지만, PARCOR 계수를 위해서는 기준패턴 2와 4가 우수하다.

표 4. LPC-cepstrum 계수의 화자확인 오차율
Table 4. Speaker verification error rate of LPC-cepstrum coefficients.

화자	화자A	화자B	화자C	화자D	화자E
C 1	0	42	22	26	16
C 2	2	18	13	10	23
C 3	0	26	3	6	7
C 4	0	14	5	18	5
C 5	5	19	11	22	6
C 6	4	18	0	1	8
C 7	10	13	20	16	6
C 8	3	17	14	0	6
C 9	0	11	10	0	5
C10	3	2	17	0	3
C11	3	12	17	1	3
C12	0	2	24	0	7
전체	0	10	0	0	10

이 실험결과에 의하면 기준패턴들이 모두 비슷한 성능을 보이고 있기 때문에, 어느 한가지 패턴이 가장 좋다고 결론지을 수는 없다. 그러나, 실제 화자인식 시스템의 구성에 있어서 기준패턴 1, 2, 3을 작성한다는 것은 복잡한 일이다. 왜냐하면 이들 패턴들은 화자인식 실험을 통하여 가장 좋은 경우를 선택하였기 때문이다. 반면에 기준패턴 4는 주어진 데이타를 단순히 평균한 것으로서 그 작성방법이 간단하다. 그럼에도 불구하고 다른 기준패턴들과 유사한 인식성능을 보이므로 실제 화자인식 시스템을 구성함에 있어서는 기준패턴 4 를 사용하는 것이 바람직하다 하겠다.

표 4 는 5명의 화자에 대하여 12차 LPC-cepstrum 계수 각각의 화자확인 오차율을 기준패턴 4를 사용하여 실험한 것이다. 즉, LPC-cepstrum 계수 하나만 사용하여 기준패턴과 시험패턴을 작성하고, 성능분석을 수행한 결과이다. C1, C2, …, C12는 각각 LPC-cepstrum의 첫번째 계수, 두번째 계수, …, 열두번째 계수를 나타내고, 마지막 항은 이들 계수를 모두 사용하여 인식한 결과이다. 이 때 기준화자 5인은 표 1 의 기준화자 5인과는 별개로 취해졌다.

이 표에서 어는 개별 계수를 사용하는 것보다는 계수 전체를 사용하는 것이 더 좋은 것을 알 수 있다. 또한 LPC-cepstrum 계수간에 성능의 차이가 심한 것을 알 수 있다. 즉, C6, C10, C12 등은 오차율이 낮은 편이나, C1, C2, C5, C7 등은 오차율이 높은 것을 볼 수 있다. 또 화자에 따라서는 12개 모두를 사용한 것보다 더 높은 인식율을 보이는 계수가 있음을 발견할 수 있다. 예를들어 화자 E의 경우 C3~C12를 별개로 사용할 때가 C1~C12 전체를 사용할 때보다 오차율이 낮은 것이다. 이것은 단일 데이타보다 복수개의 데이타가 정보량이 더 많기 때문에 더욱 좋은 인식결과를 보일 것이라는 기대와는 다르다. 그 이유는 복수개의 데이타를 결합하여 판단하는 방법이 부적합하였기 때문이라고 할 수 있겠다.

그러므로 각 계수들을 바탕으로 인식판단을 함에 있어서, 좀더 적절한 방법으로 계수들을 조합하는 판단논리가 필요함을 알 수 있다. 즉 기준패턴과 시험패턴간의 거리를 구함에 있어서, 단순한 유클리드 거리를 적용하기 보다는 각 계수간에 서로 다른 가중치를 두는 일반화 된 거리의 개념을 도입할 필요가 있는 것이다.

Ⅳ. 판단논리의 향상

화자인식의 마지막 단계에서는 시험데이타를 발음한 화자의 신원을 판단한다. 이 때 각 요소들 혹은 각 계수들간에 서로 다른 가중치를 두어 거리를 측정하는 판단논리가 필요하다는 사실을 앞장에서 확인하였다. 이와 관련하여 각 요소나 계수의 통계적 성질을 고려하는 가중치 W에 관한 연구가 이미 수행된 바 있다.[1,6,7] 즉, 화자인식을 위한 요소들의 분산을 측정하여 그 역수로 가중치를 두는 방법, F-ratio로 가중치를 두는 방법, 공분산 행렬(covariance matrix)의 역행렬로 가중치를 두는 방법 등이 시도 되었으며, 공분산 행렬의 역행렬로 가중치를 두는 방법이 일반적으로 가장 신뢰도가 있는 것으로 받아 들여지고 있다.[10]

그러나, 이러한 방법들이 최적의 가중치라고 할 수는 없다. 왜냐하면, 분산의 역수나 F-ratio로 가중치를 두는 경우에는 요소들간의 상관관계를 고려할 수 없고, 공분산 행렬의 역행렬로 가중치를 두는 경우에는 외부화자의 통계적 성질을 고려하지 못하기 때문이다. 그러므로 기준화자와 외부화자의 통계적 성질을 모두 고려해 줄 수 있고, 각 요소들의 상관관계도 아울러 고려할 수 있는 가중치가 필요하다 하겠다. 이를 위하여 본 장에서는 새로운 결정함수의 개념을 도입한 판단논리를 제안하고자 한다.

1. 거리 개념의 일반화

먼저 기준패턴과 시험패턴 사이의 유사도를 측정하기 위한 거리의 개념에 관해서 살펴보기로 하겠다. 이를 위해 그림 3 과 같은 분포를 가지는 2차원 요소벡터들을 생각해 보자. 이 그림은 화자 2명의 요소벡터들을 보인 것으로, 0은 화자 1의 데이타를 나타낸 것이고, ■은 화자 2의 데이타를 나타낸 것이다. 또, 점 x는 화자 1의 기준벡터 $\vec{r}$을 나타낸 것이다.

기준벡터 $\vec{r}$과 각각의 요소벡터 $\vec{t}$사이의 거리를 유클리드 거리, 즉

$$d(\vec{t},\vec{r}) = (\vec{t}-\vec{r})^t(\vec{t}-\vec{r}) \qquad (9)$$

와 이에 상응하는 문턱값 D_{th}를 비교하여 요소벡터 $\vec{t}$가 화자 1인지 여부를 판단한다고 하자. 먼저, 유클리드 거리를 사용하여 거리가 문턱값 $D_{th1}(=1)$보다 작으면 화자 1이라고 판단하는 경우, 이것은 그림 5에서 반경이 1인 원 C1안에 있는 벡터들만 화자1이라고 판단하는 것이 된다. 그러나, 이 때에는 화자 1에 속하는 벡터중 상당수가 화자 1이 아니라고 잘못 판단하게 된다. 이러한 경우를 피하기 위하여 문턱값을 좀더 크게하여 $D_{th2}(=4)$로 정한다면, 이 경우에는 물론 반경이 2인 원 C2안에 있는 벡터들만 화자 1이라고 판단하게 된다. 이 때 화자 1의 데이타들은 모두 옳게 판단하지만, 화자 2의 데이타가 화자 1이라고 잘못 판단되는 경우도 발생하게 된다. 그러므로 그림 3과 같은 상황에서는 원을 기준으로 화자식별을 정확히 하는 것은 불가능함을 알 수 있다. 즉, 식(9)와 같은 유클리드 거리로는 화자판단을 정확히 할 수 없는 것이다.

이러한 문제를 해결하기 위해서 유클리드 거리 대신에 일반화된 거리

$$d^W(\vec{t},\vec{r}) = (\vec{t}-\vec{r})^tW(\vec{t}-\vec{r}) \qquad (10)$$

을 도입할 필요가 생긴다. 이 식에서 행렬 W는 대칭인 양의 정치(positive definite) 행렬이다. 식(9)의 유클리드 거리를 사용하는 경우 문턱값에 해당하는 거리는 원을 형성하였으나, 식(10)의 일반화된 거리를 사용하는 경우에는 타원을 형성하게 된다. 이 때 행렬 W에 의해서 타원의 모양이 결정되며, W가 단위행렬인 경우에는 타원의 모양이 원으로 환원된다. 예를 들어, $W=\begin{bmatrix} \frac{5}{3} & -\frac{3}{5} \end{bmatrix}$인 경우에는 장축이 45° 방향에 있고, 장단축의 비가 2 : 1인 타원을 형성하게 된다. 이 때 문턱값 D_{th}를 8로 취하면, 타원 모양은 그림 3의 E와 같게 된다. 그러므로 식(10)의 일반화된 거리를 사용하면 화자 1만을 정확히 판단하는 것이 가능함을 알 수 있다.

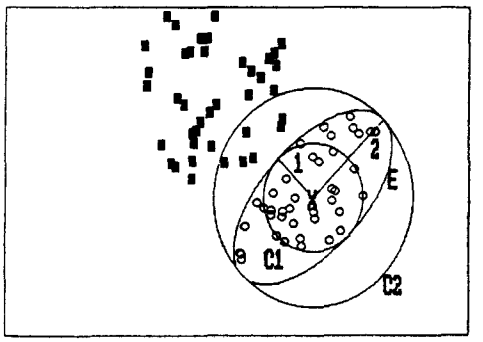


그림 3. 유클리드 거리와 일반화된 거리의 비교
Fig. 3. Comparison between Euclid distance and generalized distance.

식(10)에서 W=I인 경우 유클리드 거리에 해당하고, $W=C^{-1}$인 경우 공분산 행렬 방법을 적용한 경우에 해당한다. 이 때, 공분산 행렬은 기준화자의 분포만을 고려할 수 있기 때문에 가장 적절한 거리라고 할 수는 없다. 따라서, 본 논문에서는 외부화자의 통계적 성질도 고려해 주는 일반화된 거리 W를 구하고자 한다.

식(10)의 벡터간의 거리를 패턴간의 거리로 연장시키면, 식(7)과 같은 유클리드 거리는 거리 $D^w(T,R)$, 즉

$$D^w(T,R)=\frac{1}{N}\mathrm{tr}[(T-R)^t W(T-R)]. \qquad (11)$$

로 일반화 시킬 수 있게 된다. 이 때 W는 해당 문턱값 D_{th}와 더불어, 학습데이타로 부터 학습과정을 통해서 구해 낼 수 있다.

2. 제안된 결정함수

앞절에서 설명한 일반화 한 거리 $D^w(T,R)$을 이용하여 결정함수를

$$g(T)=-[D^w(T,R)-D_{th}] \qquad (12)$$

과 같이 정의한다. 식(11)의 거리 $D^w(T,R)$을 행렬 W의 원소에 대해서 전개하면

$$\begin{aligned} D^w(T,R)&=\frac{1}{N}\sum_{i=1}^{N}(\vec{t}_i-\vec{r}_i)^t W(\vec{t}_i-\vec{r}_i)\\ &=\frac{1}{N}\sum_{j=1}^{M}w_{jj}\Big[\sum_{i=1}^{N}(t_{ij}-r_{ij})^2\Big]\\ &+\frac{1}{N}\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}w_{jk}\Big[2\sum_{i=1}^{N}(t_{ij}-r_{ij})(t_{ik}-r_{ik})\Big] \qquad (13)\end{aligned}$$

가 된다. 표기를 간략화 하기 위하여 s_{jk}를

$$s_{jk}=\begin{cases}\frac{1}{N}\sum_{i=1}^{N}(t_{ij}-r_{ij})^2, & j=k\text{인 경우}\\ \frac{2}{N}\sum_{i=1}^{N}(t_{ij}-r_{ij})(t_{ik}-r_{ik}), & j\neq k\text{인 경우}\end{cases} \qquad (14)$$

로 정의하면, 식(11)은

$$D^w(T,R)=\sum_{j=1}^{M}w_{jj}s_{jj}+\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}w_{jk}s_{jk} \qquad (15)$$

이 된다. 그러므로 결정함수 g(T)는 결국

$$g(T)=-[\sum_{j=1}^{M}w_{jj}s_{jj}+\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}w_{jk}s_{jk}-D_{th}] \qquad (16)$$

로 표기할 수 있다.

3. 화자인식을 위한 판단 논리

미지의 시험패턴 T에 대하여 식(16)으로 정의된 결정함수값 g(T)를 구하면, 이것의 정부여부가 곧 판단논리가 되며, 따라서 식(8)은 다음과 같이 변형된다.

$g(T)\geqq 0$이면 기준화자,
$g(T)<0$이면 외부화자. (17)

앞에서 거리를 이용하여 화자를 판단했을 경우 기준패턴과 시험패턴사이의 거리가 문턱값보다 크면 외부화자라고 판단하고, 문턱값보다 작으면 기준화자라고 판단했었다. 그런데 이러한 판단은 결정함수 g(T)를 식(12)와 같이 두고 식(17)과 같은 논리를 사용하여 판단하는 것과 같다. 그러므로 결정함수를 이용한 판단논리는 앞에서 사용했던 거리를 이용한 판단을 일반화 시킨 것이라고 할 수 있다.

식(17)과 같은 논리를 적용하여 신원을 판단하기 위해서는 학습과정에서 결정함수를 구해야 한다. 거리를 이용한 판단을 사용했을 경우에는 학습과정에서 문턱값 D_{th}만 구하면 되었었다. 이 때 문턱값 D_{th}는 학습용 데이타를 바탕으로 결정되었다. 그러나, 식(16)과 같은 결정함수를 이용하여 신원을 판단하는데 있어서는 w_{jk}, $j=1,2,\cdots,M$, $k=1,2,\cdots,M$과 D_{th}가 미지수이기 때문에 이 값들을 모두 구해야만 결정함수가 완성된다. 그러므로 이 경우 학습과정에서는 문턱값 D_{th}를 구함과 동시에 행렬 W를 구할 수 있어야 하겠다. 이를 위해서는 신경회로망(neural network) 알고리듬의 하나인 perceptron방법을 이용하여, 다음과 같이 구할 수 있다.[8]

[단계 1] 행렬 W와 문턱값 D_{th}의 초기값 W^0, D_{th}^0 를 지정하고, n=0로 둔다. 이 때 W^0는 기준

화자의 공분산 행렬의 역행렬로 두고, D_{th}는 임의로 지정한다. 단, 상단 첨자 n는 반복횟수를 나타낸다.

〔단계 2〕 식(14)와 (15)를 이용하여 n번째 패턴 T^n에 대한 거리 $D^w(T^n, R)$을 계산한다.

〔단계 3〕 i) T^n가 기준화자인의 패턴의 경우 :
만일 $D^w(T^n, R) \leqq D_{th}$이면 단계 4로 넘어가고, $D^w(T^n, R) > D_{th}$이면 w_{jk}와 D_{th}를 다음과 같이 변화시킨다.

$$w_{jk}^{n+1} = w_{jk}^n + s_{jk}^n$$
$$D_{th}^{n+1} = D_{th}^n - c \qquad (18)$$

ii) T^n가 외부화자의 패턴의 경우 :
만일 $D^w(T^n, R) > D_{th}$이면 단계 4로 넘어가고, $D^w(T^n, R) \leqq D_{th}$이면 w_{jk}와 D_{th}를 다음과 같이 변화시킨다.

$$w_{jk}^{n+1} = w_{jk}^n - c\ s_{jk}^n$$
$$D_{th}^{n+1} = D_{th}^n + c \qquad (19)$$

윗 두식에서 상수 c는 변화량의 크기를 조절해 주는 인수로써, 이 알고리듬이 수렴하기 위해서는 c가 다음식과 같은 범위에 있어야 한다.

$$0 < c < \frac{2\left|\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} w_{jk}^n s_{jk}^n - D_{th}^n\right|}{1+\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}(s_{jk}^n)^2} \qquad (20)$$

〔단계 4〕 w_{jk}와 D_{th}의 변화량이 미리 정한 허용 범위내에 들어오면, 이 계산 과정을 종료한다. 그렇지 않은 경우, n=n+1로 두고 단계 2로 간다.

4. 화자인식 실험 및 결과

제안된 결정함수를 이용한 판단논리의 성능을 유클리드 거리에 의한 판단, 공분산 행렬의 역행렬로 가중치를 두는 거리를 이용한 판단 등과 비교하기 위한 화자확인 실험을 수행하였다. 이 때 학습 및 시험패턴은 Ⅲ장에서와 마찬가지 방법으로 작성하였고, 실험데이타는 8명의 기준화자에 대해서 각각 50개씩 취하고, 외부화자의 데이타 80개를 취하였다. 이 가운데 기준화자의 데이타 25개와 외부화자의 데이타 40개는 학습과정용으로 사용하고, 나머지 데이타는 시험과정에서 사용하였다.

학습과정에서는 우선 기준화자의 학습용 데이타 25개로 부터 12차의 LPC-cepstrum 계수를 추출하여 Ⅲ장에서 설명한 패턴 4와 같은 기준패턴을 작성하였다. 이 때, LPC-cepstrum 계수를 사용한 이유는 Ⅲ장의 결과로 부터 이 요소가 가장 인식성능이 우수하다는 사실을 관찰하였기 때문이다. 유클리드 거리를 이용하여 식(8)과 같은 논리로 판단하기 위해서는 학습과정에서 FR/FA 곡선으로 부터 문턱값을 정하였다. 공분산 행렬의 역행렬로 가중치를 두는 공분산방법을 이용하기 위해서는 우선 기준화자의 데이타 25개로 부터 공분산 행렬을 구하였다. 이어서 공분산 행렬의 역행렬로 가중치를 둔 거리를 적용하여 FR/FA 곡선을 그리고, 이로부터 문턱값을 정하였다. 제안된 결정함수를 이용하기 위해서는 앞절의 알고리듬을 적용하여 식(16)의 결정함수를 구하였다.

이와같은 방법으로 "통신연구실"이라는 발음을 사용하여 화자확인 실험을 수행한 결과, 표 5와 같은 인식실패율을 얻었다. 이 때 유클리드 거리를 이용한 방법과 공분산을 이용한 방법에 대해서는 식(8)을 사용하여 기준화자 여부를 판단하고, 결정함수 방법의 경우에는 식(17)에 의거하여 판단을 하였다. 유클리드 거리를 사용한 방법에 비해서 공분산 행렬을 사용한 방법의 인식율이 좋아진 것을 알 수 있다. 또 제안된 결정함수에 의한 방법은 공분산 행렬의 방법에 비할 때 4명의 화자에 대해서는 인식율이 향상되었고, 1명에 대해서만 인식율이 떨어졌다. 그러므로, 유클리드 거리를 사용한 방법보다는 공분산에 의한 방법이 더 낫고, 공분산에 의한 방법보다는 제안된 결정함수에 의한 방법이 더 우수하다고 말할 수 있겠다. 참고로, 이 결과 얻어진 공분산 행렬의 역행렬 C^{-1} 및 제안된 결정함수의 행렬 W은 각각 표 6 및 표 7과 같다. 기존의 방법들 가운데 가장 신뢰도가 높은 공분산 행렬과 비교할 때 본 논문에서 제안한 W가 대각선 성분에 있어서는 유사하나 비대각선 성분에 있어서는 차이가 많이 남을 관찰할 수가 있다.

표 5. 화자인식의 실패율
Table 5. Failure rate of speaker recognition.

화　자	화자 1	화자 2	화자 3	화자 4	화자 5
유클리드거리	3.1	3.1	3.1	10.8	13.8
공분산 방법	3.1	0.0	1.5	4.6	7.6
제안된 거리	1.5	0.0	0.0	3.1	4.6
화　자	화자 6	화자 7	화자 8	평균	분산
유클리드거리	0.0	0.0	0.0	4.2	4.0
공분산 방법	0.0	0.0	0.0	2.1	2.3
제안된 거리	0.0	1.5	0.0	1.3	1.4

표 6. 공분산 행렬의 역행렬 C^{-1}의 계수
Table 6. Coefficients of inverse covariance matrix C^{-1}.

(단위 : 10^{-2})

j\k	1	2	3	4	5	6	7	8	9	10	11	12
1	0.92	-0.07	0.11	-0.30	-0.51	0.17	0.21	0.21	0.26	0.79	0.59	0.49
2	-0.07	2.51	-0.05	-0.52	0.78	1.30	-1.56	0.96	-1.71	1.03	-2.34	3.46
3	0.11	-0.05	3.64	1.04	-0.07	-0.81	0.96	-1.75	0.57	-0.23	0.14	-3.01
4	-0.30	-0.52	1.04	4.73	-0.09	0.33	-0.46	-0.09	-0.40	-0.16	-0.01	-1.91
5	-0.51	0.78	-0.07	-0.09	5.98	-1.14	0.92	-2.08	0.39	-1.19	-0.06	-0.83
6	0.17	1.30	-0.81	0.33	-1.14	9.93	-3.16	3.69	-2.76	1.87	-2.81	5.38
7	0.21	-1.56	0.96	-0.46	0.92	-3.16	9.76	-3.96	4.27	-2.86	3.56	-6.97
8	0.21	0.96	-1.75	-0.09	-2.08	3.69	-3.96	14.34	-4.63	3.27	-0.97	4.22
9	0.26	-1.71	0.57	-0.40	0.39	-2.76	4.27	-4.63	14.86	-3.08	6.15	-5.20
10	0.79	1.03	-0.23	-0.16	-1.19	1.87	-2.86	3.27	-3.08	13.41	-2.71	8.75
11	0.59	-2.34	0.14	-0.01	-0.06	-2.81	3.56	-0.97	6.15	-2.71	17.26	-6.53
12	0.49	3.46	-3.01	-1.91	-0.83	5.38	-6.97	4.22	-5.20	8.75	-6.53	36.08

표 7. 결정함수의 가중행렬 W의 계수
Table 7. Coefficients of weighting matrix W in decision function.

(단위 : 10^{-2})

j\k	1	2	3	4	5	6	7	8	9	10	11	12
1	1.18	-0.12	-0.04	-0.47	-1.07	0.26	0.40	0.42	0.42	1.57	1.19	0.97
2	-0.12	2.79	-0.17	-0.88	1.47	2.56	-3.02	1.89	-3.42	2.28	-4.57	6.94
3	-0.04	-0.17	3.70	2.34	-0.17	-0.62	1.89	-3.46	1.19	-0.47	0.29	-5.99
4	-0.47	-0.88	2.34	4.80	-0.19	0.58	-0.86	-0.15	-0.86	-0.26	0.06	-3.80
5	-1.07	1.47	-0.17	-0.19	6.06	-2.21	1.78	-4.18	0.75	-2.32	-0.10	-1.69
6	0.26	2.56	-0.62	0.58	-2.21	9.98	-6.36	7.34	-5.53	3.76	-5.64	10.73
7	0.40	-3.02	1.89	-0.86	1.78	-6.36	9.82	-7.90	8.53	-5.71	7.15	13.91
8	0.42	1.89	-3.46	-0.15	-4.18	7.34	-7.90	14.37	-9.26	6.51	-1.93	8.44
9	0.42	-3.42	1.19	-0.86	0.75	-5.53	8.53	-9.26	14.88	-6.15	12.28	-10.41
10	1.57	2.28	-0.47	-0.26	-2.32	3.76	-5.71	6.51	-6.15	13.48	-5.35	17.47
11	1.19	-4.57	0.29	0.06	-0.10	-5.64	7.15	-1.93	12.28	-5.35	17.29	-13.05
12	0.97	6.94	-5.99	-3.80	-1.69	10.73	13.91	8.44	-10.41	17.47	-13.05	36.09

인식실험에 있어서 화자 7의 경우 제안된 결정함수 방법을 사용했을 때 인식실패율이 0%에서 1.5%로 증가한 것은 가중치를 구하는 알고리듬의 문제인 것으로 간주된다. 이것은 perceptron 알고리듬을 사용하면 데이타군들이 완전히 분리되어 있는 경우에는 오히려 최적의 가중치를 구하지 못할 수가 있기 때문이다. 즉, 이 알고리듬은 오차가 발생할 때에만 가중치를 변화시켜 주므로, 두 집단이 완전히 분리되어 있는 경우에는 오히려 최적의 상태에 도달하기 전에 이미 가중치의 변화가 중단될 수 있는 것이다. 그러므로, 이러한 상황을 감안하여 알고리듬을 개선시킨다면, 인식율은 좀더 향상될 수 있을 것이다.

V. 검토 및 결론

본 논문에서는 화자인식 시스템의 인식율 향상을 도모하기 위하여 요소의 선택 및 기준패턴의 작성문제를 고찰하였다. 또한 화자인식 실험을 수행하는 과정에서 판단논리가 인식율에 어떠한 영향을 미치는가를 아울러 검토해 보았다.

먼저 화자인식을 위한 요소들 가운데 LPC, PARCOR 계수 및 LPC-cepstrum 계수 등을 택하여 그 인식성능을 분석해 보았다. 이를 위해 위의 요소들 각각을 사용하여 화자인식시스템을 구성하고, 4가지 서로다른 기준패턴을 작성하여 화자확인 오차율을 측정하였다. 이 때 기준패턴과 시험패턴 사이의 유사도를 측정하기 위해서 유클리드 거리를 사용하였다. 그 결과 LPC, PARCOR 계수, LPC-cepstrum 계수의 화자확인 오차율은 평균 14%, 7%, 4%로 각각 나타났다. 이는 LPC-cepstrum 계수가 일반적으로 우수하다는 기존의 사실이 기준패턴에 관계없이 적용됨을 보여주며, 또한 영어로 실험된 결과가 한국어에도 적용됨을 확인해 준다. 또한 화자에 따라서도 인식율의 차이가 크다는 사실을 관찰하였는 바, 이는 화자의 발음 습관이 인식율에 큰 영향을 미칠 수 있기 때문인 것으로 사료된다. 기준패턴의 작성방법에 따라서도 인식율에 차이를 보이는데, 단순히 주어진 데이타의 평균을 취하여 작성한 기준패턴 4가 다른 기준패턴과 대등한 성능을 보였다. 그러므로, 기준패턴 4와 같은 방식으로 기준패턴을 작성함이 인식율 향상에 유리하다는 사실을 알 수 있었다. 이어서 LPC-cepstrum 계수 각각의 성능을 분석해 본 결과, 각 계수간에 성능의 차이가 큰 것을 관찰하였다. 그러므로, LPC-cepstrum 계수간의 거리를 측정하는데 있어서, 유클리드 거리 대신에 일반화된 거리를 사용할 필요가 있음을 알 수 있다.

또한 일반화된 거리 개념에 바탕을 둔 결정함수를 이용하여 새로운 판단논리를 제안하였다. 제안된 판단논리는 기준화자의 통계적 성질뿐만 아니라 외부화자의 통계적 성질까지 고려하여 화자의 신원을 판단할 수 있다는 점이 기존의 방법들과는 다르다고 할 수 있다. 유클리드 거리 및 공분산 행렬의 역행렬로 가중치를 둔 거리를 이용한 판단과 제안된 방법을 비교하기 위하여 화자인식 실험을 수행한 결과 평균 4.0%, 2.3%, 1.4%의 화자확인 실패율을 각각 얻었다. 그러므로 제안된 판단논리가 기존의 방법들 가운데에서 가장 신뢰도가 좋은 공분산 행렬의 역행렬로 가중치를 두는 방법에 비해서 더 우수함을 관찰할 수 있었다.

화자인식 시스템에서 인식율을 향상시키기 위해서는 화자인식을 위해 이미 알려진 요소들 이외의 더욱 적합한 새로운 요소의 추출이 필요하다. 이러한 연구결과에 본 논문에서 제안한 판단논리를 결합하면 더욱 높은 인식율이 가능하리라고 기대된다.

參 考 文 獻

[1] L.R. Labiner and R.W. Schafter, *Digital Processing of Speech Signals*, Prentice-Hall, New Jersey, 1978.

[2] J.T. Buck, D.K. Burton and J.E. Shore, "Text-dependent speaker recognition using vector quantization," *Proceedings, ICASSP*, pp. 381-384, 1985.

[3] 강문기, "일반화된 성도 모델에 기반을 둔 선형예측기법," 석사학위 논문, 서울대학교, 1988.

[4] M.R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. on ASSP.*, vol. ASSP-23, no. 2, pp. 167-172. April , 1975.

[5] A.E. Rosenberg and M.R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. on ASSP.*, vol. ASSP-23, pp. 254-272, April 1975.

[6] S. Furui, "Cepstrum analysis technique for automatic speaker verification," *IEEE Trans. on ASSP.*, vol. ASSP-29, pp. 254-272, April, 1981.

[7] M.R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. on ASSP.*, vol. ASSP-24, no. 4, pp. 403-409, Aug. 1976.

[8] J.T. Tou and R.C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley Publishing Company, Massachusetts, 1974.

[9] C. Myers, L.R. Rabiner and A.E. Rosenberg "Performance tradeoffs in dynamic time warping algorithm for isolated word recognition," *IEEE Trans. on ASSP.*, vol. ASSP-28, no. 6, pp. 623-635, Dec., 1980.

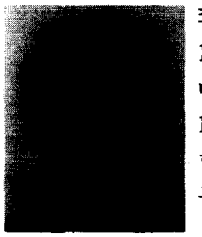
[10] M. Shridhar, N. Mohankrishnan and M.A. Sid-Ahmed "A Comparison of Distance Measures for Text-Dependent Speaker Identification," *Proc. of ICASSP 1983*, vol. 2, pp. 559-592.

[11] D. O'Shaughnessy "Speaker Recognition," *IEEE ASSP Magzine*, vol. 3, no. 4, pp. 4-17 Oct., 1986. *

著 者 紹 介

李 秉 基(正會員)

1951年 5月 12日生. 1974年 서울대학교 전자공학과 학사학위 취득. 1978年 경북대학교 대학원 전자공학과 공학석사학위 취득. 1982年 University of California, Los Angeles 공학박사 학위 취득. 1974年～1979年 해군사관학교 전자공학과 교관. 1982年～1984年 미국 Granger Associates 연구원. 1984年～1986年 미국 AT&T Bell Laboratories 연구원. 1986年～현재 서울대학교 전자공학과 조교수. 관심분야는 디지탈신호처리, 광대역통신 및 광통신, 회로이론 등임.



李 赫 宰(正會員)

1965年 2月 8日生. 1987年 서울대학교 전자공학과 학사학위 취득. 1989年 서울대학교 대학원 전자공학과 공학석사학위 취득. 주관심분야는 음성신호처리임.