대표 평균치 패턴과 가중켑스트럼을 이용한 화자인식의 성능향상에 관한 연구

정 종 순, 경연정, 최승호, 이황수
한국과학기술원 정보 및 통신공학과

A Study on the performance improvement of speaker recognition using average pattern and weighted cepstrum

Jongsoon Jung, Younjeong Kyung, Seungho Chio, Hwangsoo Lee
Dept. of Information and Communication Engineering, KAIST

요 약

본 논문은 DTW를 사용한 텍스트종속 화자확인의 성능향상에 관한 것으로, 화자인식의 근본적인 난점인 화자 정보 추출의 어려움, 사칭자의 거부, 시간 변화에 따른 인식률 저하 등을 해결하고자 하였다. 먼저, 기존의 DTW 방식을 유지하면서 DTW의 단점이라 할 수 있는 과다한 계산량과 발성 습관과 시간 변화에 따른 음성특징을 개선하기 위하여 기준 패턴에 통계적 의미를 도입한 대표 평균치 패턴을 사용하였다. 그리고 화자간의 변별력을 극대화하기 위하여 가중 켑스트럼을 제안하였다. 가중 켑스트럼은 화자별로 유용한 켑스트럼 차수를 구하여, 그 차수에 가중치를 두는 것으로 본 실험에서는 F-ratio를 사용하여 구하였다. 실험결과 대표 평균치 패턴과 F-ratio를 사용할 경우 인식률이 각각 약 3 ~ 4% 향상 되었다.

1. 서론

사회의 정보화가 급속히 진행됨에 따라 통신망을 통한 사용자의 대규모 데이타베이스에 대한 접근과 정보의 검색, 갱신, 수정이 빈번해지고 있다. 이에 따라 정보의 보안 문제가 심각해지고, 특정 지역의 출입 통제를 위한 보안 시스템의 필요성이 증대되고 있다. 따라서 사용자의 본인 여부를 판단하는 개인 확인 수단이 필수적이다. 그러나 종래의 개인 확인 수단으로 널리 쓰이는 카드, 도장, 신분증 등은 도난이나 위조에 취약한 실정이다. 특히 정보의 접근이 전화나 통신망등을 이용하여 원격지에서 이루어지는 경우, 개인 확인은 더욱 어려워진다. 이에 비해 화자인식은 음성에 포함되어 있는 화자정보를 추출하여 개인을 확인하는 기술로서 사칭자(imposter)에 대한 대처, 처리시간, 원격지 확인 등 여러 측면에서 가장 효과적인 기술 중 하나이다. 또한, 음성을 확인 매체로 이용하기 때문에 사용이 편리하고 응용 분야가 넓다는 장점이 있다 [1,2].

본 논문에서는 우선, 기존의 dynamic time warping(DTW)의 단점이라 할 수 있는 계산량과 음성특징에 따른 인식률 저하를 개선하기 위하여 수개의 기준 패턴들로부터 한개의 대표 평균치 패턴을 DTW 방식과 통계적 개념을 도입하여 구하였다. 두번째로는 F-ratio를 사용하여 파라미터를 개방함으로써 화자간의 변별력을 향상시키고자 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 본 논문에서 기술한 화자인식 시스템을 개략적으로 살펴본다. 3절과 4절에서는 대표 평균치 패턴 구성의 알고리즘과 제안한 가중치 켑스트럼을 설명한다. 그리고 5절은 실험결과를 보이고 6절에서 결론을 맺는다.

2. 화자인식 시스템

화자인식은 인식 방법에 따라 패턴정합법(pattern matching)인 DTW, 신경회로망, 벡터양자화, hidden Markov model 등 크게 4가지 방식으로 구분한다. 그리고 인식 대상에 따라 발성화자를 구분해내는 화자식별(speaker identification)과 본인 여부를 판단하는 화자확인(speaker verification)으로 나눌 수 있으며, 인식에 사용하는 문장의 종속 여부에 따라 텍스트종속(text dependent)형과 텍스트독립(text independent)형으로 나눈다 [2]. 화자인식 시스템의 전체적인 흐름도를 그림 2.1에 보였다.

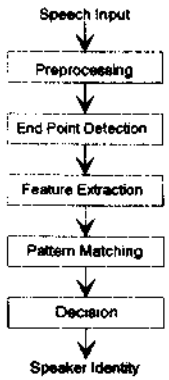


그림 2. 1 화자 인식 시스템의 전체적인 흐름도

그림에서 보듯이 음성파형은 sampling, A/D변환 등 전처리 과정과 음성구간 검출 과정을 거친다. 그리고 검출된 음성신호는 특징추출 과정을 거치며, 추출된 음성 파라미터열은 DTW에 의한 패턴 정합을 통과하면서 화자인식 여부가 결정된다.

2.1. 음성구간 검출

음성구간 검출의 정확성 여부는 음성인식의 성능에 큰 영향을 미치기 때문에 정확한 검출이 요구된다. 따라서 본 논문에서는 단구간 에너지와 영교차율(zero crossing rate)의 문턱치(threshold)를 adaptive thresholding method를 사용하여 조절하는 방법을 사용하였다. 즉 단구간 에너지는 입력된 데이타만을 이용하여 이 값이 큰 부분은 음성구간으로 작은 구간은 묵음구간으로 결정하는 방법이다. 그러나, 이러한 방법은 무성자음과 같이 에너지가 작은 부분이나 배경잡음이 큰 경우에는 음성과 묵음을 구별하기가 어렵다. 따라서 이를 보완하기 위하여 영교차율이 사용하였다. 그리고 어떤 단어에는 pause가 존재하는데 위와 같은 방법을 사용하면 pause의 전에서 끝점이 검출될 수 있다. 따라서 본 논문에서는 이러한 오류를 개선하기 위하여 끝점이 검출된 이후에 40 프레임내에 다시 음성의 시작점이 검출되면 pause가 있는 음성으로 간주하고 다시 끝점 검출을 한다. 이때 pause 검색구간을 0.4 초로 하였다. 끝점을 찾아도 pause 검색으로 인식시간이 길어지기 때문이다. 지금까지 설명한 끝점 검출 알고리즘을 통하여 얻은 음성 데이타는 다음절에서의 특징추출 과정에 전달된다 [3,4].

2.2 음성특징 추출

우선, 30 ms의 Hamming window를 사용하여 한 프레임을 구성하였으며, 10 ms씩 옮긴다. 그리고 1 - 0.98z-1의 preemphesis를 거친 데이타로부터 LPC 계수를 구한다. 이 LPC 계수로부터 cepstrum을 구하고, 귀의 비선형적인 특성을 고려해 cepstrum을 mel-scale로 warping시켜 얻은 mel-cepstrum을 특징 파라미터로 사용하였다. 그림 2.2는 이러한 특징추출 과정을 나타낸다 [5,6].

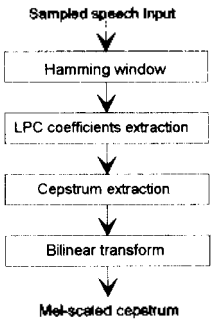


그림 2. 2 특징 추출 과정

2.3 DTW를 이용한 패턴정합

같은 화자가 같은 단어를 발성할 경우라도, 발성음의 길이는 매시간마다 비선형적으로 전개와 수축하면서 변화한다. DTW는 입력음성과 기준 패턴을 정합하기 위하여 시간축상에서 비선형적으로 왜곡시키는 과정이다. 이 과정은 dynamic programming 기술을 이용한다. 다음 그림 2.3은 DTW 알고리즘의 일반적인 패턴정합을 나타낸 것이다 [7].

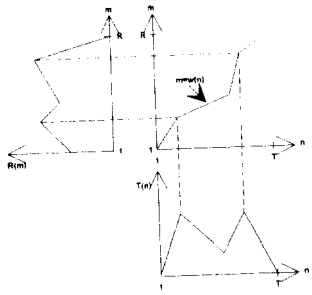


그림 2.3 DTW 알고리즘에 의한 입력패턴과 기준패턴의 비선형 패턴정합

패턴정합의 기본 아이디어는 각 프레임별로 거리를 계산하여 가장 최소거리를 갖도록 하는 것이다. 즉, 입력 패턴 T(n)에 대해 거리 $D = \min_{w(n)}[\sum_{n=1}^{T} d(T(n),R(w(n)))]$ 가 최소가 되는 기준 패턴 R(w(n))을 찾는 것이다. 이러한 패턴정합법을 이용한 화자인식 시스템은 다수 보고되었다 [8,9,10].

3. 대표 평균치 패턴 구성의 알고리즘

정확한 화자인식을 위해서는 유효한 특징 파라미터와 특징 파라미터간의 유사성을 평가하는 척도의 선택이 필요하다. 특징 파라미터는 화자의 개인성 정보를 충분히 표현할 수 있어야 하며, 화자내의 변이는 충분히 수용하되 화자간의 변이는 최대화하는 것이 분별력을 높이는데 바람직하다. 기존의 DTW 방식은 동적 프로그래밍으로 인해 계산량이 많고, 화자내의 변이를 수용할 수 있는 기준패턴의 작성이 어려웠다. 본 논문에서는 수개의 기준 패턴들로부터 통계적인 방법을 사용하여 한개의 기준패턴(기준 평균치 패턴)을 구하였으며, 이를 이용하여 인식할 경우 계산량이 크게 줄어들었다. 일반적인 N개의 기준패턴을 사용한 경우와 대표 평균치 패턴을 사용한 경우의 성능을 비교하기 위하여 다음과 같이 화자 확인 시스템을 구성하였다. 시스템의 대략적인 블럭도는 그림 3.1과 같다.

본 시스템의 입력은 customer의 음성과 imposter의 음성이다. 우선 customer에 해당하는 각 사람에 대하여 2절에서 언급한 알고리즘을 사용하여 음성의 특징 파라미터와 mel-cepstrum을 추출한다. 이 특징 파라미터열은 대표 평균치 패턴과의 거리계산에 의해 customer의 가부를 결정한다. 자세한 설명은 다음에서 언급한다.

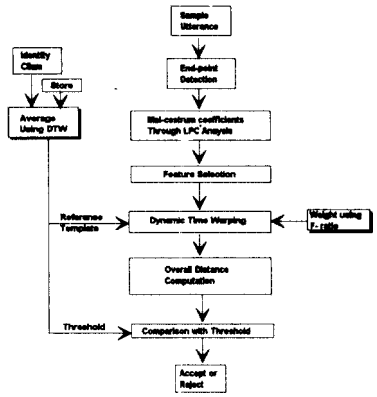


그림 3. 1 대표 평균치 패턴을 사용한 시스템의 개략적인 블럭도

3.1 문턱치

모든 입력 음성과 기준 패턴간의 거리를 구하여, customer를 주장하는 사람을 받아들일 것인지 거부할 것인지를 결정하는 문턱치와 비교한다. customer인 화자에 대한 거부, 즉 false rejection 확률과 imposter를 받아들이는 false acceptance 확률 등 2 가지 가능한 에러가 있다. 일반적으로, 문턱치는 이들 에러에 대한 비중에 따라 선택한다. 따라서 이 문턱치를 어떻게 설정하느냐도 중요하다. 만약 문턱치를 낮게 설정하면 false rejection 확률은 높아지나 false acceptance 확률은 낮아진다. 반대로, 이 문턱치를 높게 잡으면 false rejection 확률은 낮아지나 false acceptance 확률은 높아진다. 따라서 필요에 따라 false acceptance 확률을 낮추어 사용 할 수도 있고, false rejection 확률을 낮추어 사용할 수도 있다. 그러나 일반적인 경우 그림 3.2 에 보인 것과 같이 false acceptance 확률과 false rejection 확률의 중간에서 문턱치를 결정한다.

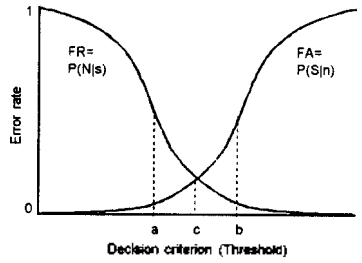


그림 3. 2 에러율과 문턱치 사이의 관계

따라서 본 논문에서도 여러가지 실험을 토대로 이들값을 균등하게 맞추어서 문턱치로 사용하였다.

3.2 대표 평균치 패턴

일반적으로 DTW를 이용한 패턴정합 방식은 기준패턴을 여러개 사용하여 테스트 패턴과 각각 모두를 비교한다. 본 논문에서는 기준패턴에 해당하는 사람의 음성을 시간에 따라 받아서 training 과정을 거쳐 하나의 대표 평균치 패턴을 구한다. 이는 기준패턴 정보의 갱신과 구성이 시스템에 또 다른 중요한 요소이기 때문이다.

만일 training 패턴이 N개인 경우, 대표 평균치 패턴은 다음과 같이 구한다. 우선, DTW에 의해 첫번째 training 패턴의 각 프레임별로 이에 해당하는 두번째 training 패턴의 프레임과의 평균을 구한다. 그리고 이 결과 패턴과 세번째 training 패턴을 같은 방식으로 평균을 구한다. 이와 같이 하여 N번째 패턴까지 수행하면 최종 패턴(대표 평균치 패턴)을 구한다.

시간에 따라 음성을 받을 때는 시스템에 따라 일정한 기간을 두어야한다. 만약 어떤 한 customer에 음성을 오랜 기간 동안에 걸쳐 받아 평균을 구하여 기준패턴으로 사용하면, false acceptance 확률이 증가하게 되어 화자인식률를 저하시키는 원인이 된다. 그리고 짧은 기간 동안 발성된 음성을 사용하면 반대로 false rejection 확률이 높아져 이것 또한 인식률을 저하시키는 원인이 된다. 본 논문에서는 6개의 training 패턴으로부터 대표 평균치 패턴을 구하였다.

4. F-ratio를 이용한 켑스트럼 가중치 설정

화자간의 변별력을 극대화하기 위하여 가중 켑스트럼 파라미터를 제안한다. 가중켑스트럼은 화자별로 특히 유용한 켑스트럼 차수가 어떤 것인지 측정한 뒤, 그 차수에 가중치를 두는 것으로 본 논문에서의 실험 결과 매우 유용한 파라미터임을 확인하였다.

본 논문에서는 가중치로 F-ratio 값을 사용하였다. F-ratio는 특징파라미터의 유용성 척도로 주로 사용되는 것으로 화자내의 variance로 화자 간의 variance를 나눈 값이다. F-ratio의 정의는 식(4.1)과 같다[8].

$$F-ratio = \frac{variance\ of\ speaker\ means}{mean\ intraspeaker\ variance} \qquad (4.1)$$

좋은 특징 파라미터는 화자간의 변이는 크고 화자 내의 변이는 작은 것이다. 즉, F-ratio 값이 클수록 좋은 파라미터라 볼 수 있다. 본 논문에서는 이러한 F-ratio의 특징에 착안하여 이를 각 켑스트럼 차수별로 측정하여 실제 어떤 켑스트럼 차수들이 화자간의 변별력을 높이는지를 알아내고 이에 가중치를 두었다. 위 정의에 따라 본 논문에서 사용한 F-ratio는 다음과 같다. 식(4.2)는 일반적인 F-ratio 값을 구하기 위해 사용한 식이고 식(4.3)은 이를 각 화자별로 적용한 것이다. 식(4.3)의 결과는 켑스트럼의 가중치 $W(i)$로 사용된다.

$$F-ratio = \frac{Var(E(C_{ij})_j)_{\text{전체 화자}}}{E(Var(C_{ij})_j)_{\text{전체 화자}}} \qquad (4.2)$$

$$i = 1,\ldots Order \quad j = 1,\ldots No.\ of\ speakers$$

$$W(i) = \frac{Var(E(C_{ij}))_{\text{전체 화자}}}{(Var(C_{ij}))_{\text{각 화자별}}} \qquad (4.3)$$

$$i = 1,\ldots Order \quad j = No.\ of\ each\ speakerS$$

만약, 위에서 구한 F-ratio의 켑스트럼 차수에 대한 분포가 일정하거나 거의 변화가 없다면 화자의 개인성 정보는 켑스트럼 차수에 대해 거의 동일하게 분포되어 있다는 것을 의미한다. 그러나 차수에 따라 F-ratio 값의 변화가 심하게 나타난다면 이는 특정 차수의 켑스트럼 값이 화자인식에 매우 유용하게 적용된다는 것을 알려준다. 그림 4.1은 식(4.2)에서 구한 일반적인 F-ratio값을 보이고 있다. 그림에서 볼 수 있듯이 F-ratio 값은 켑스트럼 차수별로 그 값의 차이가 심하게 나타난다. 이는 특정 차수의 값이 화자식별에 더 유용하게 사용될 수 있음을 보인다. 그림 4.2는 식(4.3)에 따라 F-ratio를 화자별로 구분하여 구한 결과를 보이고 있다. 본 논문에서의 시스템은 사용권한이 있는 화자의 수가 4명인 경우로, 그림 4.2에서 각 화자에 대해 F-ratio 값의 분포가 달리 나타남을 알 수 있다. 이는 위에서 구한 가중치가 화자의 변별력을 높이는 좋은 요소가 될 수 있음을 보인다.

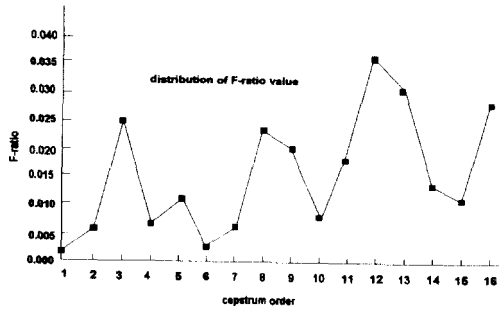
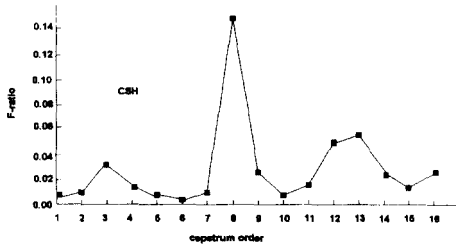
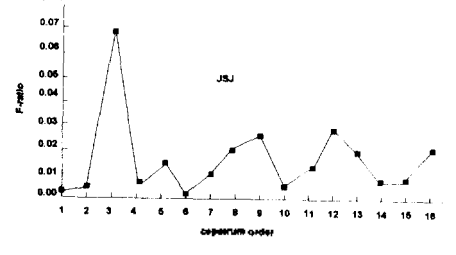


그림 4. 1 켑스트럼 차수에 의한 F-ratio 값의 분포

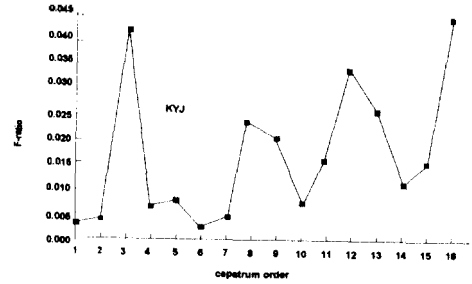
a. 화자 1



b. 화자 2



c. 화자 3



d. 화자 4

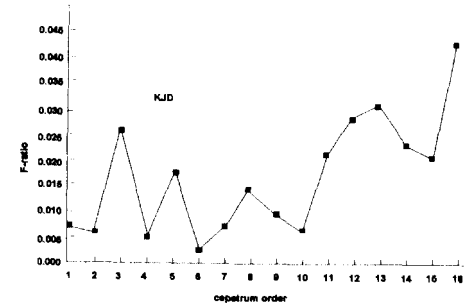


그림 4. 2 화자별 켑스트럼 차수에 대한 F-ratio 값의 분포

5. 인식 실험 및 결과

5.1. 데이타베이스 구성

본 논문에서는 사용한 음성데이타는 일반 실험실 환경에서 성인 남녀 10명이 발성한 음성으로 구성하였다. 이들은 20대 중반에서 30대 초반의 남녀로 customer 4명(여자:2명, 남자:2명)과 imposter 6명(여자:3명, 남자:3명)이다. Customer는 기준패턴으로 사용하기 위해서 일주일간에 걸쳐서 아침, 점심, 저녁으로 나누어 본인의 이름과 다른 customer의 이름을 3번씩 발성하였다. 그리고 imposter는 각 customer의 이름을 위와 같은 방법으로 3번씩 발성하였다. 테스트 패턴은 한달 후 같은 방법으로 customer는 10번, imposter는 3번을 발성하여 얻었다.

- ▶ 녹음 환경 : 일반 실험실 환경
- ▶ A/D 변환 : 8kHz sampling, 16 bit linear PCM, ASPI사의 ELF DSP보드 사용 [11].
- ▶ 음성 DB 내용 : 3음절에 해당하는 이름
- ▶ 전체 발음 횟수 : customer와 imposter 200회

5.2. 대표 평균치 패턴 사용 결과

앞에서 언급한 대표 평균치 패턴을 사용한 결과를 그림 5.1에 보였다. 이 그림은 4명의 customer에 대한 인식률을 N개의 패턴을 사용할 경우와 비교하여 보인 것이다. 그림에서 imposter로 사용하는 발성음은 기간적으로 골고루 선택하였다.

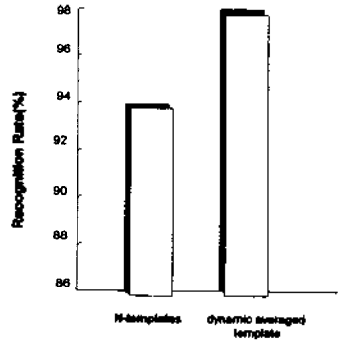


그림 5. 1 대표 평균치 패턴 사용시 결과

그림에서 보는 바와 같이 대표 평균치 패턴을 사용할 경우의 인식률이 3 - 4% 향상되었다. 그리고 <표 1>에는 (a) N개의 패턴을 사용할 경우와 (b) 대표 평균치 패턴을 사용할 경우로 나누어 각 customer의 에러개수를 비교해 보였다. 이 결과로부터 (b)시스템의 성능이 (a)보다 우수한 것을 알 수 있다.

표 1. 화자별 False acceptance와 False rejection 개수
(a) 대표 평균치 패턴 사용시

(개)	customer1 (csh)	customer2 (jsj)	customer3 (kjd)	customer4 (kyj)
FA	1	0	1	0
FR	0	0	0	0

(b) 일반적인 N개의 기준패턴 사용시

(개)	customer1 (csh)	customer2 (jsj)	customer3 (kjd)	customer4 (kyj)
FA	2	1	1	2
FR	0	0	0	0

5.3 가중 켑스트럼 사용 결과

앞에서 살펴 본 화자별 F-ratio 값을 화자확인 시스템의 켑스트럼에 대한 가중치로 사용하여 실험을 행하였다. F-ratio 값을 적용하지 않은 경우 인식률은 94%이다. 이에 반하여 가중 켑스트럼을 사용한 경우 본 시스템의 인식률은 98%로 향상되었다. 이는 특히 약 한달 전의 데이타를 기준패턴으로 한 경우이므로 시간 변화에 따른 일반적인 인식률 저하가 가중 켑스트럼을 사용함으로써 해결될 수 있음을 보인다. 그림 5.2는 가중 켑스트럼을 사용한 실험결과이다.

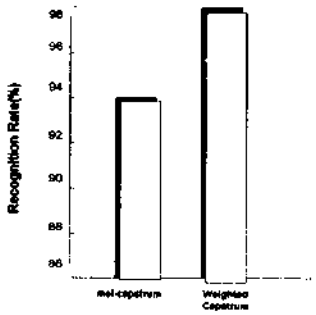


그림 5 2 가중 켑스트럼을 사용한 결과

6. 결론 및 향후 연구 방향

본 논문은 DTW를 사용한 텍스트 종속 화자확인에 관한 것으로, 기존의 DTW 방식을 유지하면서 이것의 단점이라 할 수 있는 과다한 계산량과 발성 습관에 관한 음성왜곡을 개선하기 위한 것이다. 우선 수개의 기준 패턴들로부터 한개의 대표 평균치 패턴을 DTW 방식과 통계적 개념을 도입하여 구하였다. 그리고 켑스트럼에 가중치를 두어 화자간의 변별력을 향상시켰다. 위의 두 방식 각각을 실험한 결과, 일반적인 방식과 비교하여 약 3 - 4%의 인식률 향상을 보였다.

이후, 대표 평균치 패턴으로부터 구한 가중켑스트럼에 의한 인식성능 비교와 기준패턴 구성시 기간에 따른 인식성능 차이에 대한 연구를 하고자 한다.

[참고 문헌]

[1] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, Inc. , 1992.

[2] L. R. Rabiner & R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978.

[3] 윤성진, "적은 학습자료 환경 하에서 화자인식 시스템의 성능향상에 관한 연구," KAIST 석사논문 , pp.1-14, 1993.

[4] 구명완외, "실기간 음성 끝점검출 알고리즘," 제 5 회 신호처리 합동학술회 논문집, 제 5권 1호, pp. 11-14, 1992.

[5] 이세웅, 최승호 외, "이산 HMM을 이용한 실시간 음성인식 다이얼링 시스템 개발," 한국음향학회지, 제 13권 1E호 , pp.89-95, 1994.

[6] G. D.Forney, "The Viterbi Algorithm," *Proc. of the IEEE*, vol. 61, NO. 3, pp. 268-278, March 1973.

[7] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, Springer-Verlag, New York, N.Y., U.S.A, 1978.

[8] S. Furui & A. E. Rosenberg, "Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech," in Proc. ICASSP, vol. 5, pp. 1060-1062.

[9] H.Ney and R.Gierloff, "Speaker Recognition Using a Feature Weighting Technique," in Proc. ICASSP'82, pp.1645-1648, 1982.

[10] J.M.Naik and G.R.Doddington, "High Performance Speaker Verification Usign Principal Spectral Components,", in Proc. ICASSP'86, pp.881-884, 1986.

[11] Atlanta Signal Processors Inc., "ELF DSP Plateform Introduction Manual," 1992.

[12] Douglas O'Shaughnessy, "Speaker Recognition", IEEE Trans. on ASSP Magazine, pp.4-17, 1986.

[13] Atlanta Signal Processors Inc., "ELF DSP Plateform Introduction Manual," 1992.

[14] Hiroaki Sakoe & Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on ASSP, vol. 26, No. 1, pp.91-97, Feb. 1978.

[15] H. Ney, "The Use of a One-Stage Dynamic programming Algorithm for Connected Word Recognition," IEEE Trans. on ASSP, vol.61, No.3, pp.268-271, April 1989.