화자인식 기술

이 황 수
한국과학기술원 정보 및 통신공학과

Speaker Recognition Technique

HwangSoo Lee
Dept. of Information and Communication Engineering, KAIST

요 약

본 논문에서는 화자인식 기술의 필요성과 방법에 관하여 간단히 서술하고 있다. 통신망을 통한 개인의 정보 검색이 증가하면서 간편하고 정확한 화자인식 기술의 필요성이 증대되고 있다. 종래의 개인 확인 수단인 신분증, 도장, 서명 등은 원거리에서 통신망을 이용하여 정보를 이용하고자 할 때 적용되지 못하며 부가적인 장치가 요구된다. 이에 반해 음성을 이용하여 사용자를 확인하는 화자인식 기술은 별다른 부가적인 장치가 필요하지 않고 편리하게 이용할 수 있는 장점을 지닌다.

1. 서론

정보화 사회가 발달함에 따라 통신망 등을 통한 사용자의 데이타베이스나 시스템에 대한 정보 처리 요구와 접근이 급격히 증가하고 있으며, 아울러 이에 따른 정보의 보안 문제가 심각해지고 있다. 한편, 점점 개방화되는 사회 분위기로 인하여 특정 지역의 출입 통제를 위한 보안 시스템의 필요성이 증대되고 있다. 이러한 이유로 사용자의 본인여부를 판단하는 개인 확인 수단이 필수적이나, 종래의 개인 확인 수단으로 널리 쓰이는 카드, 도장, 서명, 신분증 등은 도난이나 위조의 문제점을 안고 있으며, 특히 정보의 접근이 전화나 통신망 등을 이용한 원거리에서 이루어질 경우 개인에 대한 확인은 더욱 어려워진다[1].

화자인식은 음성에 포함되어 있는 화자정보를 추출하여 개인을 확인하는 기술로 전화망을 통한 서비스가 증대되고 있는 현대 사회에 가장 효과적인 기술중 하나이다. 음성을 이용한 화자인식은 사용이 편리하고 별다른 장치의 부가가 필요치 않아 그 응용범위가 넓은 장점을 지닌다.

화자인식 기술의 필요성은 크게 다음 2가지로 나누어 생각할 수 있다.

1) 보안성

전술한 바와 같이 화자인식의 가장 큰 필요성은 보안에 있다. 음성을 이용한 화자인식 기술에 의한 사용자 제한은 전화망 등 원거리 정보 이용뿐 아니라 컴퓨터 소프트웨어에도 적용 가능하다. 현재 상용화되어 있는 여러 소프트웨어는 보안을 위해 암호입력과 같은 장치를 해 두었으나 이는 타인에게 공개될 위험이 있고 때로는 사용자가 암호를 잊어버리는 경우도 있게 된다. 이를 사용자의 음성을 이용한 화자인식 기술로 대치할 경우 사용이 간편하고 타인에게 암호가 도용될 위험도 없게 된다. 음성에는, 어떤 말인지를 나타내는 "언어정보"와 함께 화자가 누구인지를 알 수 있는 "개인성"이 있다. 화자 인식은 이 중 개인성을 나타내는 정보를 보아 사용권한이 있는 사람인지를 판별한다.

2) 사용자별 정보 제공의 차별화

통신망을 통한 정보 제공과 데이타베이스 검색의 경우 사용자의 사용레벨에 따라 그 이용 한도와 범위제한이 필요하게 된다. 음성을 이용한 화자인식 기술은 이러한 처리를 용이하게 해 준다. 여러 사용자에 대한 화자식별을 통해 각 개인의 사용레벨을 보아 차등화한 정보 제공을 할 수 있도록 한다. 이는 사용자에 따라 서로 다른 서비스를 제공해야 할 경우에도 적용 가능한 기술이다. 이처럼 화자인식 기술은 다중 사용자에 대한 개별적인 정보제공과 서비스가 가능하도록 하는 필수적인 기술이다.

화자인식 기술은 다음과 같은 응용이 가능하다.

(1) 음성을 이용한 "성문" 감지 보안장치 제작
(2) 전화망을 이용한 자동응답시스템(ARS)의 보안장치로 사용
(3) 멀티미디어 환경에서의 사용자 편의 제공 소프트웨어 개발을 위한 요소 기술
(4) 통신망 상에서 홈-뱅킹 등의 서비스 제공을 위한 개인식별 보안장치
(5) 개인 전자 비서 시스템 개발에 필수적인 요소기술 확보
(6) 다중 사용자에 대한 데이타베이스의 차별 제공을 위한 수단

(7) 각종 전자제품이나 교육용 소프트웨어에 응용가능한 요소 기술 확보 등.

본 논문에서는 화자인식 기술의 필요성을 간략히 살펴본 서론에 이어 2장에서 화자인식 시스템의 개요를 보고 3,4장에서 실제 시스템의 모델과 파라미터를 본다. 5장 결론에서는 연구동향과 더불어 앞으로의 연구방향에 대하여 살펴본다.

2. 자동 화자인식(ASR:Automatic Speaker Recognition)의 개요

화자 인식(Speaker Recognition)에는 크게 화자 확인(Speaker Verification)과 화자 식별(Speaker Identification)의 두 분야가 있다. 이 두 분야는 모두 N명에 대한 참조패턴(reference pattern) 데이타베이스를 갖고 있다는 점에서는 서로 같지만, 어떠한 일을 하느냐에 따라 다음과 같이 구별된다. 화자 확인은 입력신호로서 음성 신호와 그 음성 신호의 화자에 대한 Identity가 같이 주어지며, 시스템은 그 Identity에 해당되는 화자에 대한 참조패턴(reference pattern)과 입력된 음성 신호가 일치하는 지를 검사하여 일치 또는 불일치라는 판정을 내리게 된다. 즉, i번째 화자에 대해 다음과 같은 결정이 내려진다.

i번째 화자 확인 : If $d(T(x), Pi(x)) \leq$ 문턱치

i번째 화자 거부 : If $d(T(x), Pi(x)) >$ 문턱치

$T(x)$: 검사할 입력 패턴, $Pi(x)$: i번째 화자에 대한 참조패턴, $d(A,B)$: A, B의 거리

반면 화자 식별은 입력 신호로서 들어온 음성 신호가 검토 대상인 N명 중 누구에게 해당되는 지를 찾아 해당 화자가 누구인지를 알려주어야 한다[2]. 즉, 다음과 같은 기준으로 i번째 화자가 선택된다.

i번 화자 선택 : If $d(T(x), Pi(x)) < d(T(x), Pj(x))$ for all speakers $i \neq j$

$T(x)$: 검사할 입력 패턴, $Pi(x)$: i번째 화자에 대한 참조패턴, $d(A,B)$: A, B의 거리

이 때 입력 신호가 검토 대상인 N명 중 어느 누구에게도 해당되지 않을 경우도 고려해야 한다[3]. 이처럼 대상 외의 화자가 발성할 가능성을 고려한 경우를 open-set 화자 구성이라 하고, 발성할 화자가 반드시 N명 내에 있다는 가정 하에 구현된 시스템은 close-set 화자 구성이라 한다.

화자 인식의 일반적인 블럭도는 그림 1에 보이고 있다. 음성이 입력되면 그의 특징(feature)을 뽑아 참조패턴 또는 검사패턴으로 사용하며 두 패턴간의 유사도를 측정하여 그 결과 값에 따라 화자 인식을 행한다.

화자 확인 시스템의 에러는 false rejection과 false acceptance의 두 가지 에러로 나눌 수 있다[3]. false rejection은 옳은 사용자에 대해 잘못 거부하는 것이고, false acceptance는 반대로 사용권한이 없는 화자를 올바른 사용자로 오인식 하는 것이다. 이 두 가지 에러는 서로 trade-off 관계에 있다. 즉, false acceptance 에러율을 줄이기 위해 유사성 비교의 문턱치를 낮추어 주면 false rejection 에러율이 증가하고 반대로 false rejection 에러율을 줄이기 위해 문턱치를 너무 높이면 false acceptance 에러가 증가하게 된다. 그림 2에 두 에러 사이의 관계를 보이고 있다.

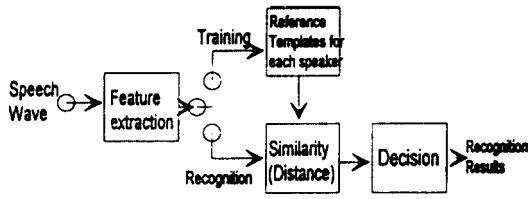
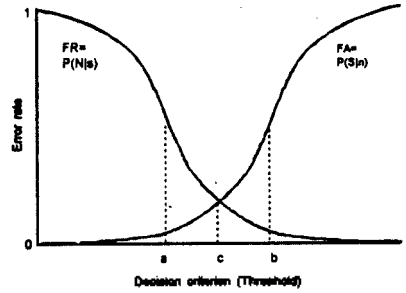


그림 1. 화자 인식 시스템의 블럭도



FR(False Rejection), FA(False Acceptance), N:거부, S:확인, n:사칭자, s:옳은 화자

그림 2. 에러율과 문턱치 결정의 관계

그러므로 시스템 구성 시 이러한 관계를 잘 고려하여 적절한 문턱치를 가질 수 있도록 신중한 실험이 필요하다. 아울러 이는 시스템의 사용 용도에 따라 특정 에러를 줄이는 방향으로 결정할 수 있다. 예를 들어 철저한 보안이 요구되는 경우엔 false rejection 에러율이 좀 높아지더라도 false acceptance 에러율을 특정 기준 이하로 떨어뜨려 놓아야 할 것이다.

3. 화자인식 방법

화자인식 방법은 그 알고리즘에 따라 다음 4가지 모델로 크게 나눌 수 있다.

3.1 패턴정합법에 의한 화자인식

패턴정합법(template matching)은 DTW(Dynamic Time Warping) 등을 사용하여 입력된 신호의 패턴을 미리 정해진 참조패턴과 비교하여 유사성을 판단하는 기법이다. 이 방법은 길이가 다른 두 패턴을 비선형적으로 정합하는 방법으로 최적화(global optimization)를 수행하는 특성이 있다. 즉, DTW는 발

성시 시간축에 대해 변이가 많음을 고려하여 참조패턴과 비선형적으로 비교할 수 있게 해준다. 이 방법은 사용어휘가 제한되는 단점이 있으나 비교적 간단한 시스템으로 좋은 성능을 얻을 수 있는 것으로 알려져 있다. 다음 그림 3은 DTW 알고리즘의 일반적인 패턴정합을 나타낸 것이다[4].

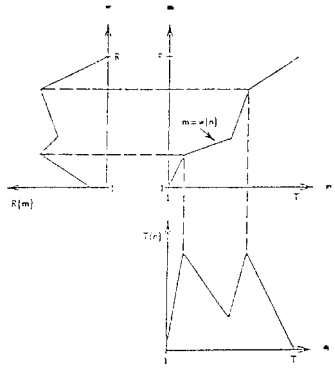


그림 3. DTW 알고리즘에 의한 입력 패턴과 참조패턴의 비선형 패턴정합

패턴정합의 기본 아이디어는 각 프레임별로 거리를 계산하여 가장 최소거리를 갖도록 하는 것이다. 즉, 입력패턴 T(n)에 대해 $D = \min_{w(n)}[\sum_{n=1}^{T} d(T(n),R(w(n)))]$ 가 최소가 되는 참조패턴 R(w(n))을 찾는 것이 알고리즘의 목적이다. 이러한 패턴정합법을 이용한 화자인식 시스템으로 [5]-[8]이 보고되어 있고 문종속 화자인식의 경우 비교적 간단한 시스템으로 좋은 성능을 갖는다.

3.2 신경회로망을 이용한 화자인식

신경회로망(NN:neural network)을 이용하는 방법은 각 화자별로 신경회로망을 구성하고 화자간의 분별력을 갖도록 학습을 수행한다. 이 방법은 신경 회로망이 가지는 고도의 병렬계산 능력, 적응성, 분별 학습에 의한 높은 화자 식별율과 사용어휘에 제한을 받지 않는 특성을 갖는다. 그러나 back-propagation 알고리즘을 이용한 지도학습(supervised learning) 모델의 경우, 학습 데이타들이 모두 학습되어질 때까지 반복해서 연결강도의 값을 변경해야 하며, 특히 새로운 화자의 추가시 전체 데이타를 모두 재학습시켜야 하는 단점이 있다. 신경회로망을 이용한 화자인식 시스템으로 [9]-[14]가 보고되어 있고 한정된 화자수인 경우 좋은 결과를 내는 것으로 알려져 있다.

3.3 벡터양자화에 의한 방법

벡터 양자화(VQ:vector quantization)에 의한 화자인식은 패턴과 양자화 코드북(codebook) 사이의 거리로 두 패턴 사이의 유사성을 판별하는 방법이다. 화자별 양자화 코드북은 화자의 특징 파라미터의 분포를 겹치지 않는 몇 개의 영역으로 나누어 대표하고, 입력 신호의 패턴에 대한 시간축의 정합없이 패턴 공간상에서의 분포 특성만을 비교하여 인식을 수행한다. 벡터 양자화 방법은 사용 어휘에 제한을 받지 않으면서 화자 모델의 크기가 작아 실용 시스템에 유용하나, 많은 학습 자료가 필요하고 음성의 동적인 변화 특성을 이용하지 못한다는 단점이 있다. 벡터 양자화를 이용한 시스템은 [15]-[19]가 보고되어 있으나 최근들어 점점 다른 방법과 함께 쓰이는 보조적인 수단으로 많이 쓰이는 추세이다.

3.4 HMM을 이용한 방법

HMM(Hidden Markov Model)을 이용한 방법은 현재 음성인식 분야에서 널리 사용되고 있는 확률모델을 화자인식에 적용하는 방법이다. HMM은 학습 기능을 이용하여 화자 내의 변이를 흡수할 수 있으며, 입력 패턴의 비선형 정합을 수행하는 특성이 있다. 이 방법은 모델의 구성 형태에 따라 텍스트 종속이나 텍스트 독립 시스템 구현이 가능하고 다수의 화자를 대상으로 한 화자 인식에도 효과적이라 알려져 있다. 그러나 다수의 화자를 위한 모델을 만들기 위해서는 많은 training 데이타가 필요한 단점이 있다[1]. HMM을 이용한 화자인식 시스템은 [20]-[24]가 있고 최근까지도 연구가 활발히 진행되고 있다.

4. 화자인식 파라미터

특징벡터는 화자의 개인성 정보를 충분히 표현할 수 있어야 하며, 화자간 유사성을 평가하는 척도는 화자 내의 변이를 수용하면서도 화자 사이의 변이는 최대화할 수 있는 것이어야 한다[6].

좋은 특징은 사용자내(intra speaker)에서는 그 변이가 작고, 사용자간(inter speaker)의 벡터 사이에서는 그 변이가 크다. 즉, 그림 4와 같은 분포를 갖는 특징벡터가 좋은 벡터라 할 수 있다.

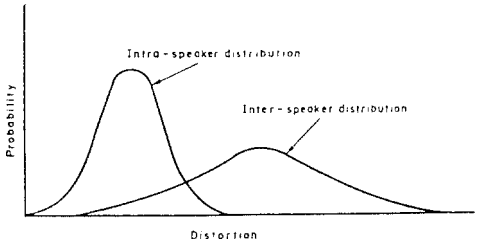


그림 4. 좋은 특징벡터의 분포

좋은 특징벡터가 갖추어야 할 조건은 ①화자간 정보를 효과적으로 나타낼 수 있어야 하고, ②측정이 쉬워야 하고, ③시간에 대해 안정해야 하고, ④음성신호에서 자연적이고 반복적으로 일어나야 하며, ⑤발성환경이 바뀌어도 영향이 적어야 하고, ⑥흉내를 허용하지 않아야 한다[25]. 이러한 조건을 모두

만족하는 파라미터는 아직 발견되지 않았고, 일반적으로 많이 사용되는 특징에는 다음과 같은 것들이 있다. MFCC(Mel frequency cepstral coefficients)가 대표적으로 많이 쓰이는 것이고 이 외에 △-MFCC, PLP, △-PLP 등이 최근들어 많이 연구되고 있다[26]. 아울러 주변 환경 소음에 강한 특징벡터에 관한 연구가 많이 진행된다[26]-[32]. 화자인식 모델이 실험실 환경하에서 우수한 성능을 갖더라도 실제 사용되는 환경에서 제 성능을 나타내지 못하면 그 의의성이 없어진다. 최근 들어 이러한 소음환경을 고려한 여러 특징벡터와 알고리즘 개발이 활발히 진행되고 있다. 한 예로 RASTA Processing이나 Bandpass Liftering 등이 유용한 것으로 나타나 있다[26],[29]. 요즘들어 각광받고 있는 RASTA Processing은 다음의 필터를 갖는 처리 알고리즘이다.

$$H(z) = \frac{M(z)}{A(z)} = \frac{Z^4(0.2+0.1Z^{-1}-0.1Z^{-3}-0.2Z^{-4})}{(1-0.94Z^{-1})}$$

이는 주변 잡음에 좋은 성능을 나타낸다고 보고되어있다[18],[26],[29],[33],[34].

앞에서 살펴본 바와 같이 화자인식 시스템에 여러 가지 특징벡터가 사용될 수 있고 실제 여러 가지가 연구되어 있다. 그러면 실제로 구성하는 화자인식 시스템에 어떤 특징 벡터를 선택할 것인가 하는 문제가 생기게 된다. 이를 해결하기 위해서는 특징의 효용성을 나타내는 척도를 정의해야 한다. 널리 사용되는 효용성 측정 척도로 F-ratio가 있다[6]. F-ratio는 다음과 같이 정의된다.

$$F = \frac{\text{variance of speaker means}}{\text{mean intraspeaker variance}}$$

즉, 전체 화자들의 평균치의 변이 값을 화자 각각의 변이의 평균으로 나눈 것으로 정의된다. 이는 그림4의 화자내 편차로 화자간의 편차를 나눈 효과를 갖는데 화자내 편차는 작을 수록 좋고 화자간의 편차는 클 수록 좋다. 그러므로 F-ratio 값이 크면 좋은 특징이라 말할 수 있고 값이 작으면 성능이 좋지 못한 특징으로 생각할 수 있다. F-ratio는 좋은 특징을 선택하는데 보다 나쁜 특징을 가려내는 목적으로 많이 사용된다[3].

5. 결론

본 논문에서는 화자인식 기술의 필요성과 방법에 관하여 간략하게 살펴보았다. 화자인식 기술에 대한 연구는 전화망이 설치된 이후 본격화되었고, 특히 컴퓨터를 사용한 자동 화자인식(ASR:Automatic Speaker Recognition) 기술은 1976년 Atal의 연구를 시작으로 활발해졌으며 부분적으로 상용화되기 시작했다. 그러나 외국에 비해 국내의 화자인식에 대한 본격적인 연구는 아직 없었다. 보고된 연구 결과는 [1],[2],[35]가 있으나, 실제 상용화된 예는 없다. 앞에서 살펴본 바와 같이 화자인식 기술은 그 응용범위가 넓어 연구가치가 매우 큰 분야이다. 그러므로 국내에서도 우리말에 적절한 화자인식 알고리즘과 특징에 대한 연구가 이루어져야 한다.

화자인식 시스템의 성능을 향상시키기 위해서는 앞에서 살펴본 적절한 모델과 특징벡터의 선택이 중요하다. 아울러 화자간의 개인성은 개개인의 성도의 특성 등 해부학적인 측면에서도 차이가 나지만, 나름대로의 발성습관등도 큰 영향을 미친다. 지금까지 연구되어온 대부분의 파라미터는 성도 특성등의 검출에 관한 것으로 발성습관에 관한 연구도 행해져야 하겠다.

참고문헌

[1] 윤성진, "적은 학습자료 환경하에서 화자인식 시스템의 성능향상에 관한 연구," 석사학위 논문, MSC 923314, 한국과학기술원, 1994.

[2] 임창헌, "모음인식과 벡터 양자화를 이용한 화자인식 시스템에 대한 연구," 석사학위 논문, MEE 86366, 한국과학기술원, 1988.

[3] D. O'Shaughnessy, "Speaker Recognition," IEEE ASSP Magazine, pp.4-17, Oct. 1986.

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

[5] S. Furui and A. E. Rosenberg, "Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech," in Proc. ICASSP'80, Vol. 5, pp.1060-1062, 1980.

[6] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Dekker, in Proc. ICASSP'82, 1992.

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

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

[9] Y.Bennani and P.Gallinari, "On The Use Of TDNN-Extracted Features Information In Talker Identification," in Proc. ICASSP'91, pp.385-388, 1991.

[10] L.Rudasi and S.A.Zahorian, "Text-Independent Talker Identification With Neural Networks," in Proc. ICASSP'91, pp.389-392, 1991.

[11] J.Oglesby and J.S.Mason, "Radial Basis Function Networks for Speaker Recognition," in Proc. ICASSP'91, pp.393-396, 1991.

[12] H.Hattori, "Text-Independent Speaker Recognition Using Neural Networks," in Proc. ICASSP'92, Vol.2, pp.153-156, 1992.

[13] J.M.Naik and D.M.Lubensky, "A Hybrid HMM-MLP Speaker Verification Algorithm for Telephone Speech," in Proc. ICASSP'94, Vol.1, pp.153-156, 1994.

[14] K.R.Farrell and R.J.Mammone, "Speaker Identification Using Neural Tree Networks," in Proc. ICASSP'94, Vol.1, pp.165-168, 1994.

[15] J.T.Buck, et al., "Text-Dependant Speaker Recognition

Using Vector Quantization," in Proc. ICASSP'85, pp.391-394, 1985.

[16] A.E.Rosenberg and F.K.Soong, "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes," in Proc. ICASSP'86, pp.873-876, 1986.

[17] f.K.Soong et al., "A Vector Quantization Approach to Speaker Recognition," in Proc. ICASSP'85, pp.387-390, 1985.

[18] T.Matsui and S.Furui, "A Text-Independent Speaker Recognition method Robust Against Utterance Variations," in Proc. ICASSP'91, pp.377-380, 1991.

[19] T.Matsui and S.Furui, "Comparison of Text-independent Speaker Recognition Methods Using VQ-distortion and Discrete/Continuous HMMs," in Proc. ICASSP'92, Vol.2, pp.157-160, 1992.

[20] E.Rosenberg et al., "Connected Word Talker Verification Using Whole Word Hidden Markov Models," In Proc. ICASSP'91, pp.381-384, 1991.

[21] M.J.Carey et al., "A Speaker Verfication System Using Alpha-Nets," in Proc. ICASSP'91, pp.397-400, 1991.

[22] J.J.Webb and E.L.Rissanen, "Speaker Identification Experiments Using HMMs," in Proc. ICASSP'93, Vol.2, pp.387-390, 1993.

[23] M.E.Frosyth, M.A.Jack, "Discriminating Semi-Continuous HMM for Speaker Verification," in Proc. ICASSP'94, Vol.1, pp.313-316, 1994.

[24] Chi-Shi Liu et al., "Speaker Recognition Based on Minimum Error Discriminative Training," in Proc. ICASSP'94, Vol.1, pp.325-328, 1994.

[25] 배명진, "화자식별을 위한 디지탈 음성신호처리," '88 음성신호처리워크샵 논문집, pp.47-55, 1988.

[26] J.P.Openshaw et al., "A Comparison of Composite features under Degraded Speech in Speaker Recognition," in Proc. ICASSP'93, Vol.2, pp.371-374, 1993.

[27] R.C.Rose et al., "Robust Speaker Identification in Noisy Environments Using Noise Adaptive Speaker Models," in Proc. ICASSP'91, pp.401-404, 1991.

[28] D.A.Reynolds and R.C.Rose, "An Integrated Speech-background Model for Robust Speaker Identification," in Proc. ICASSP'92, Vol.2, pp.185-188, 1992.

[29] Y.Kao et al., "Robustness Study of Free-Text Speaker Identification and Verification," in Proc. ICASSP'93, Vol.2, pp.379-382, 1993.

[30] K.T.Assaleh and R.J.Mammone, "Robust Cepstral Features for Speaker Identification," in Proc. ICASSP'94, Vol.1, pp.129-132, 1994.

[31] H.Gish et al., "Robust, Segmental Method for Text Independent Speaker Identification," in Proc. ICASSP'94, Vol.1, pp.145-148, 1994.

[32] R.Ricart et al., "Speaker Recognition in Tactical Communication," in Proc. ICASSP'94, Vol.1, pp.329-332, 1994.

[33] H.Hermansky at al., "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," in Proc. of EUROSPEECH'91, pp.1367-1370, 1991.

[34] J.Koehler at al., "Integrating RASTA-PLP Into Speech Recognition," in Proc. of ICASSP'94, pp.I-421 - I-424, 1994.

[35] 김형래 외, "주파수 에너지를 이용한 텍스트 독립 화자인식에 관한 연구," 제 11회 음성통신 및 신호처리 워크샵 논문집, pp.235-240, 1994.