論文94-31B-10-9

모음 검출을 통한 텍스트 독립 화자인식에 관한 연구

(A Study on the Text-Independent Speaker Recognition from the Vowel Extraction)

金 에 녹*, 卜 赫 圭**, 金 炯 來**

(Enoch Kim, Hyeok Kyu Bok and Hyung Lae Kim)

要 約

본 논문에서는 모음 검출을 통하여 미리 등록한 단어가 아닌 경우에도 화자를 인식할 수 있도록 특징 파라메터를 개발하고, 실용화가 가능하도록 처리 방법을 간략화한 텍스트 독립 화자인식 연구를 진행하였다. 이를 위해서, 화자가 발성한 음성에서 모음을 검출하여 화자인식에 사용하는 방법을 제안하였으며, 인식은 각 화자가 발성한 음성 신호에서 모음을 검출한 다음, 검출된 모음의 29 채널의 주파수 에너지를 퍼지값으로 표현한 후, 퍼지 추론을 적용하여 수행하였다.

실험을 위해 모음 검출 알고리듬을 개발하였으며, 화자인식의 특징 파라메터로 29 채널 주파수 에너지를 제안하였는데, 별도의 코드북 없이 사용이 가능하고, 기존의 파라메터에 비해 인식율이 높으면서도 구성 및 계산이 간단한 특징이 있다. 실험 결과, 미리 작성된 표준패턴과 동일한 단어를 사용한 텍스트 의존 화자인식 실험은 95.5 % 인식율을 보였고, 표준패턴과 다른 종류의 단어를 사용한 텍스트 독립 화자인식 실험은 94.2 % 인식율을 보이고 있다.

Abstract

In this thesis, we perform the experiment of speaker recognition by identifying vowels in the pronounciation of each speaker. In detail, we extract the vowels from the pronounciation of each speaker first. From it, we check the frequency energy of 29 channels. After changing these into fuzzy values, we employ the fuzzy inference to recognize the speaker by text-dependent and text-independent methods.

For this experiment, an algorithm of extracting vowels is developed, and newly introduced parameter is the frequency energy of the 29 channels computed from the extracted vowels. It shows the features of each speakers better than existing parameters. The advanced point of this parameter is to use the reference pattern only without the help of any codebook.

As a result, text-dependent method showed about 95.5 % rate of recognition, and text-independent method showed about 94.2 % rate of recognition.

*正會員, 蓮菴工業專門大學 電子計算科
(Dept. of Com.Sci., Yonam Junior College of Eng.)

**正會員, 建國大學校 電子工學科
(Dept. of Elec.Eng., Kon Kuk Univ.)

接受日字 : 1994年 4月 4日

Ⅰ. 서 론

기계에 의해 음성으로 화자를 자동 파악하는 화자 인식 방법은 화자확인 (speaker verification) [1] 과 화자식별 (speaker identification) [2] 로 나눌 수 있으며, 화자확인은 특정인이라고 자칭하는 인식 대상이 본인인지 여부를 판정하는 것이며, 화자식별은 발화자가 등록된 화자중에서 누구인지를 판단하는 것이다. 이때 사용되는 화자 음성의 언어적 내용이 그룹내의 화자로 미리 등록한 언어적 내용과 동일한 경우는 텍스트 의존(text-dependent) 화자 인식, 두개가 서로 다를 때는 텍스트 독립 (text-independent) 화자인식이다. [3]

대표적인 화자인식 연구는, 1963년 Bell 연구소의 Pruzansky [4] 가 10명의 화자, 10종의 단어로 시간 평균 스펙트럼을 사용해서 화자식별 실험을 하였고, 1974년 Bell 연구소의 Atal [5,2] 은 10명의 화자가 발성한 짧은 문장으로 화자식별 및 확인 실험을 하여 텍스트 독립 사용시 93% 결과를 보고하였으며, 1976년에는 LPC 켑스트럼(cepstrum) 계수가 화자인식에 우수함을 보고하였다. 1976년 Sambur [6] 는 21명의 화자가 같은 문장을 6번 반복 발음한 고음질의 음성 데이타를 직교 변화된 인자로써 연구하였다. 1974년 Furui [7], 1977년 Markel, Oshika 그리고 Gray [8] 은 시간 경과에 따른 특징량 변화에 대해서 연구하였다.

1981년 Furui [9,10] 는 LPC 켑스트럼에 기초한 화자확인 실험에서 정적(static) 인 특성뿐만 아니라 동적(dynamic)인 특성도 중요하다고 생각하여, 음성의 통계학적인 특성과 DP 매칭(Dynamic Program matching)을 비교하였으며, 이 방법으로 구한 거리를 서로 더해서 실험하여 좋은 결과를 얻었다. 또한, 1981년 미국의 Helms [11] 는 15명의 화자가 발음한 256 단어들을 선형 예측 벡터 코드북을 사용해서 실험한 결과 88% 인식율을 보고하였다. 1985년 Buck [12] 등은 벡터 양자화를 사용한 텍스트 의존 실험의 결과를 발표하였으며, 1988년 Soong과 Rosenberg [13] 는 벡터 양자화에 기초한 화자인식 실험에서 순시(瞬時) 스펙트럼 정보가 과도(過渡) 스펙트럼 정보보다 화자에 관한 정보를 잘 나타내며, 또한 두가지를 함께 사용하여서 인식율을 개선시킨 결과를 발표하였다. 그러나 벡터 양자화 방법이 정보를 압축하여 계산량은 줄일 수 있으나, 양 끝이 유사한 단어들에 대해서는 잘못된 결과를 보일 수 있으므로, 이런 문제점들을 개선한 행렬 양자화(Matrix Quantization) 기법이 제안되었다. 1987년 Burton [14] 은 행렬 양자화 기법으로 텍스트 의존 화자확인 실험을 하였으며, 1990년 Juang과 Soong [15] 그리고 1993년 Chen [16] 등은 화자인식에 있어서 벡터 양자화보다 행렬 양자화 방법이 우수함을 발표하였다. 그러나 아직은 제한된 단어나 문장을 사용한 텍스트 의존 화자인식이 주로 다루어지며, 새로운 알고리듬의 제시가 부족한 현실이다.

본 논문에서, 실험 데이타들은 데이타 전체를 그대로 사용하지 않고, 8 비트의 A/D 컨버터를 통하여 들어온 음성 데이타를 내용에 구분없이 프레임 단위로 분석한 후, 대수 에너지, 영교차율, 정규화된 상관 계수를 이용하여 모음을 검출하여 사용한다. 인식 실험에서는 화자의 음성파에 포함된 개인차와 특징 파라메터의 경시적 변화의 영향을 해결하기 위하여 퍼지 이론을 도입하고, 검출된 모음의 29 채널 주파수 에너지를 각각 소속도 함수로 표현하여 퍼지화 패턴으로 작성한다. 여기서 확신도는 퍼지 연산 max-min 에 의해 구하며, 이 패턴들 사이의 확신도를 구하여 화자를 인식하는 퍼지 추론에 의한 화자인식 방법을 제안하고자 한다.

Ⅱ. 전처리 및 음성신호 분석

실험을 위해 구성한 화자인식 시스템은 마이크를 통해 입력된 음성을 곧바로 화자인식에 적용치 않고, 왜곡을 보상하기 위하여 윈도우를 거친 음성을 사용하여 끝점추출과 이어 모음 검출을 수행한 후, 각 어휘에서 얻어진 모음에서 각 화자의 특징을 나타낼 수 있는 파라메터를 구하여 인식실험을 하였다. 끝점추출을 위해서 대수 에너지, 영교차율, 자기상관계수등의 파라메터를 사용 하였고, 각 단어에서 검출한 모음을 사용해서 주파수 대역을 29차로 나누어 각 채널의 주파수 에너지를 구하여 인식실험을 수행하였다.

음성의 주파수적 분포는 0.02 ~ 20KHz의 범위내에 있는 것으로 알려져 있고, 일상 쓰이는 말의 에너지 범위는 4KHz 이내에 집중되어 있다. Nyquist 율에 따라, 샘플링 주파수는 음성 신호의 최대 주파수의 2배 이상을 사용하면 신호에서 왜곡없이 좋은 신호 성분을 얻을 수 있다.

화자인식 실험을 위한 전처리 과정에서 4KHz의 저역 필터를 사용하여 고주파 성분을 제거한 뒤, 10KHz로 샘플링을 하였으며, 분해 정도가 8 Bit인 A/D 변환기를 사용하여 음성을 디지탈화 하였다. 먼저, 입력된 음성을 프레임 단위로 나누어 분석을 해야 하므로, 여기서 발생되는 데이타의 불연속으로 인한 주파수 성분의 왜곡을 보상하기 위한 방법으로 해밍

윈도우를 사용하였다. 또한 화자의 특징을 적절하게 내포하고 있는 음성의 구간을 검출하기 위하여, 분석된 프레임 속에서 모음을 검출하여 화자인식 시스템에 적용시켰다. 모음은 대부분 유성음으로 각 화자의 성도 특성을 잘 나타내고 있다. [17] 따라서 실험을 위해 기본 모음을 미리 발음한 뒤, 그 모음을 가지고 분석한 계수들을 이용하여 모음 검출에 적용하였다.

1. 모음 검출

한국어의 모음은 발음하는 음성마다 하나 이상을 포함하고 있으므로, 모음을 택하여 화자인식에 사용하는 방법을 제안하였다. 모음은 주로 유성음으로 보통 단독으로 발음되는 경우와 초성 또는 종성에 자음을 동반하는 경우로 분류할 수 있다. 따라서, 정확한 모음을 검출하기 위하여, 한국어의 단모음인 /아/, /에/, /이/, /오/, /우/, /어/를 미리 발음한 후, 각 화자별로 모음의 특징을 찾아 실제 음성에 적용하여 모음 검출 알고리듬을 구성하였다. 이때 모음 검출을 위해 사용한 파라메터들은 에너지, 영교차율, 정규화된 자기상관계수 이다.

음성 신호 분석은 프레임 단위로 처리하기 때문에, 12.8 ms 를 한 프레임으로 간주하여 6.4 ms 씩 겹쳐 음성 신호를 처리하였으며, 모음의 시작점은 전체 7 프레임이 임계 조건을 만족하면 현재의 프레임에서 -6 번째 프레임을 모음의 시작점으로 정하였다. 또한 연속해서 임계치를 만족하지 않을 경우에는, 임계치 Counter를 0 으로 재설정하여 다시 프레임 Counter를 증가시키는 방법으로 정확한 모음 검출을 시도하였다.

2. 모음 검출 알고리듬

화자인식 실험에 사용하는 특징 파라메터들은 음성 신호 안에 포함된 모음을 검출해서 사용하게 되는데, 이를 위해서 우선 정확한 모음 검출이 이루어져야 한다.

모음은 자음과 달리 에너지가 높고, 주기적인 성분을 포함하므로 영교차율이 자음에 비해서 작은 점등을 고려한다면 영교차율과 대수 에너지 만으로 모음 분류가 가능 하리라고 생각 하지만, 실험의 정확성을 위해서 정규화된 자기상관계수를 포함시켰다. 정규화된 자기상관계수는 실험에 의하면 모음일 경우 0.899 이하의 값을 가지므로, 1에 가까운 경우에는 잡음이나 무성음 또는 자음으로 간주하였다.

모음 검출에 사용된 알고리듬은 먼저 발음한 단모음을 분석하여 가상의 모음 시작 점의 임계점을 설정한 뒤, 실제 음성에 적용하여 임계값을 변화시키는 방법을 택한 후 적당하다고 생각되는 임계값을 다른 화자의 음성 신호에 적용하여 모음을 검출하도록 하였다.

$$IZCR = 1/N \left(\Sigma TZCR\ [i] \ * \ \delta \right)$$
$$IENG = 1/N \left(\Sigma TENG\ [i] \ * \ \delta \right)$$
$$IACR = 1/N \left(\Sigma TACR\ [i] \ * \ \delta \right)$$
$$ICEP = 1/N \left(\Sigma TCEP\ [i] \ * \ \delta \right)$$

여기서, IZCR : 임계 영교차율
IENG : 임계 에너지
IACR : 임계 정규화된 자기상관계수
ICEP : 임계 캡스트럼 계수

여기서 N는 처음 설정된 프레임에서 부터 임계조건을 만족하지 않는 프레임의 갯수를 나타내고 있고, δ는 선택된 임의의 임계치에 대한 변화율을 비율로서 나타낸 값으로 0.1씩 증가 시키는 값이다.

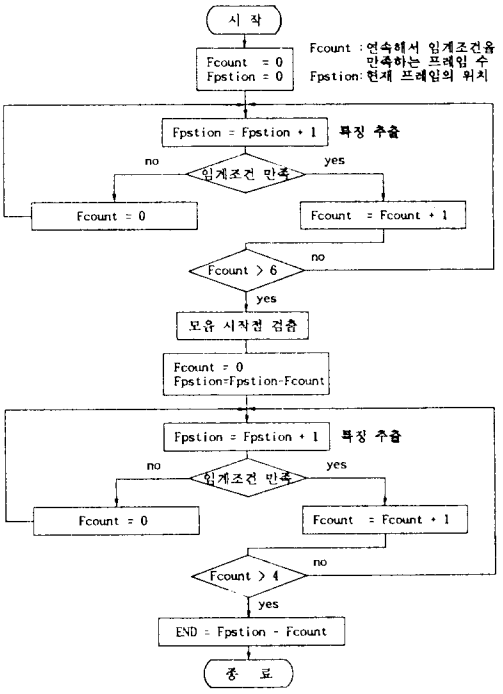

그림 1. 모음 검출 흐름도
Fig. 1. Block diagram for detecting of a vowel.

〈모음 시작점 추출〉

IF 현재 프레임이 임계조건 만족
  임계 Counter 1 증가
  IF Counter = 7 현재 프레임 모음 시작

ELSE 다음 프레임의 파라메터 계산
ELSE Counter 초기화

여기서 검출된 모음의 시작점에서 다시 모음의 끝점을 검출하기 위하여, 위에서 정의된 임계조건을 다시 한번 사용한다.

〈 모음 끝점 추출 〉
IF 현재 프레임이 임계조건 불만족
임계 Counter 1 증가
IF Counter = 5 현재 프레임 모음 끝
ELSE 다음 프레임의 파라메터 계산
ELSE Counter 초기화

3. 포만트 주파수 측정

모음은 성대가 주기적으로 개폐하는 것에 따라 생긴 진동파가 음원이 되고, 구강(口腔)과 비강(鼻腔)의 형태에 의해 결정되는 공조 특성에 의해서 변형되며, 모음의 종류에 따라 특정 주파수 성분이 강조된다. 이 모음을 특징짓는 우세한 주파수 성분이 포만트 주파수이다. 보통 모음에는 몇개의 포만트 주파수가 있으며, 이 포만트는 발성자, 성별, 연령등에 따라 큰 폭으로 변동한다. [17] 따라서, 화자의 특징을 표현하는 파라메터의 하나로 포만트 주파수의 전력 스펙트럼을 변형시켜서 사용하고자 한다.

포만트를 측정하는 방법은 켑스트럼 분석에 의한 방법과 선형예측에 의한 방법이 있는데, 선형예측에 의한 포만트 추적은 유성음에 대하여 정확한 포만트 주파수를 얻을 수 있고, 최소 복잡성, 최소 계산시간, 그리고 최고의 포만트 추정 정밀도 등의 장점이 있다. [18,19] 선형예측에 의한 포만트 주파수 추출법에는 root-solving 법과 peak-picking 법이 있는데, root-solving법의 계산이 복잡하고 시간이 많이 소요되기 때문에, 실험에서 포만트 주파수 추출은 peak-picking법을 사용하였다. 또한 포물선 보간법으로 구한 값을 첨두 위치로 사용하며, 포물선은 다음 식의 형태이다.

$$y(\lambda) = a\lambda^2 + b\lambda + c \qquad (1)$$

만약 이산적 스펙트럼 위치가 η_p에 위치한다면, 보간해서 생성된 주파수는,

$$\hat{F} = (\eta_p + \lambda_p) f_s \ / \ (2N) \qquad (2)$$

이며, 이때 f_s = 1 / T, zero 인덱스에 대한 첨두 위치 λ_p = -b / 2a 이다.

또한, 대역폭은 다음 식을 만족한다.

$$\hat{B} = \frac{-\{b^2 - 4a\,[c - 0.5y\,(\lambda p)]\}^{1/2} f_s}{aN} \qquad (3)$$

Ⅲ. 퍼지 추론에 의한 화자인식

1. 퍼지 추론의 화자인식 적용

1) 생성규칙

화자를 퍼지 추론에 의해 인식할 때, 인식을 위한 생성규칙이 필요하다. 이 생성규칙은 전제에 따른 결론의 형태로서 다음과 같이 주어진다. [20]

IF 전제 THEN 결론

본 연구에서의 생성규칙은, 각 화자가 발성한 음성 신호에서 검출한 모음으로부터 구한 각 채널의 주파수 에너지를 특징량으로 사용하여서, 다음과 같이 생성규칙을 만든다.

IF 음성 신호내 모음의
각 채널 주파수 에너지 F_i가 퍼지값 X^{Fi}를 갖는다면,
THEN
음성 신호는 화자 "S"이다.

화자인식을 하기 위해서는 퍼지 추론의 생성규칙을 이용하여 특징 파라메터에 대하여 소속도 함수를 구하고 퍼지집합을 생성시켜야 한다. 따라서 각 화자가 발성한 모음에 대해서 각 채널의 주파수 에너지에 대한 퍼지화 값을 할당하고, 주파수 에너지에 대한 퍼지집합의 전체집합 Ue이 생성되면, 이것을 이용하여 표준패턴(reference pattern)과 시험패턴(test pattern)에 대하여 소속도 함수를 구하고 퍼지집합을 생성한다.

2) 입력 패턴의 퍼지화

화자인식을 위해 구한 특징들은 그 음성을 나타내는 절대적인 것은 아니며, 이 특징량은 동일인이 같은 조건하에서 발성하여도 그때마다 조금씩 다르게 된다. 이러한 음성의 변동을 해결하기 위해 퍼지화 패턴으로 표현한다. 각 화자에 대한 음성 신호의 특징량을 각 모음에서 추출한 주파수 에너지로 한다. 이때, 주파수 에너지를 퍼지값으로 나타내기 위하여 주파수를 29 채널로 나누어 각각의 중심 주파수에 해당하는 에너지 만을 나타낼 수 있도록 한다. 여기서

구해지는 주파수 에너지의 존재 범위는 -40 dB로 제한할 수 있으므로, 1 dB 마다 퍼지값을 주어 40 개의 퍼지값으로 대응시킨다. 주파수 에너지가 -39 dB 이하가 되면, 그 값에 관계없이 퍼지값을 40으로 하였으며, 0 dB 이상이 되면 마찬가지로 퍼지값을 1로 한다.

2. 텍스트 독립 화자인식 알고리듬

1) 주파수 에너지 특징량의 퍼지화

그림 3.1 (a)는 표준패턴 모음 "아", (b)는 시험패턴 모음 "아"의 주파수 에너지의 예이다. 그림에 전력 스펙트럼을 dB 단위로 표현하기 위하여 식 (4)을 사용한다. 이 식에서, $T_F(i)$ 는 각 프레임을 29차 디지탈 필터를 사용하여 중심 주파수에 대한 에너지 값을 구한 뒤, 전체 n 개의 프레임에 대하여 각 채널별로 주파수 에너지 값을 누적하여 구하며, 다시 각 채널별로 구한 29개의 $T_F(i)$ 값을 모두 더하여 T_{FE}을 구한다.

$$L(z) = 10 \log_{10} | T_F(i) / T_{FE} | \ [dB] \qquad (4)$$

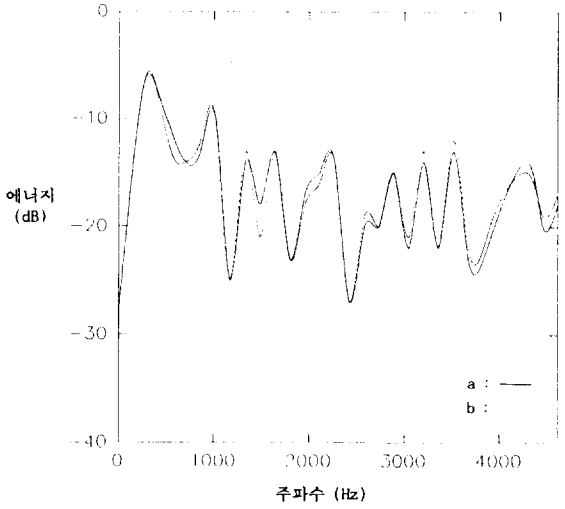

그림 2. (a) 표준패턴 모음 "아"의 주파수 에너지
(b) 시험패턴 모음 "아"의 주파수 에너지

Fig. 2. (a) Frequency energy of "아" vowel for reference pattern.
(b) Frequency energy of "아" vowel for test pattern.

표 1은 그림 2(a), (b)의 주파수 에너지의 특징량을 퍼지화 패턴으로 표현한 것이다. 표준패턴을 위해서 사용한 주파수 에너지를 퍼지값으로 표시한 것과 시험패턴을 위해서 사용한 다른 주파수 에너지를 퍼지값으로 표시한 것을 표 1에 함께 나타내었다. 이와

표 1. 주파수 에너지에 의한 특징량의 퍼지화 패턴의 예

Table 1. Example of fuzzified pattern for feature of frequency energy.

중심 주파수[Hz] \ 퍼지값	표준 패턴		시험 패턴	
	에너지 [dB]	퍼지값	에너지 [dB]	퍼지값
제 1 채널 : 0	-28	29	-28	29
제 2 채널 : 234	-08	9	-08	9
제 3 채널 : 390	-07	8	-06	7
제 4 채널 : 546	-11	12	-13	14
⋮	⋮	⋮	⋮	⋮
제 15 채널 : 2265	-14	15	-14	15
제 16 채널 : 2421	-27	28	-26	27
제 17 채널 : 2578	-20	21	-20	21
제 18 채널 : 2734	-20	21	-20	21
⋮	⋮	⋮	⋮	⋮
제 26 채널 : 3984	-19	20	-18	19
제 27 채널 : 4296	-14	15	-14	15
제 28 채널 : 4453	-20	21	-19	20
제 29 채널 : 4609	-18	19	-17	18

같이 표현되는 퍼지화 패턴을 좀 더 명확하게 설명하기 위해, 퍼지화 패턴을 퍼지값과 소속도 함수 값과의 관계로 나타낸다. 이때 소속도 함수 값은 퍼지화 패턴의 의미를 어느 정도 포함하고 있는가를 나타내는 것으로 1.0 에서 0.0 사이의 값을 가지며, 값이 1.0 인 경우에는 퍼지화 패턴의 의미를 완전히 포함하는 것이며, 0.0 일 때는 완전히 포함하지 않는 것을 의미한다. 이 예에서 0.0 과 1.0 사이의 소속도 함수 값을 중심 퍼지값에서 부터 벗어나는 정도에 따라서, 0.1 씩 감소하도록 하였다.

2) 주파수 에너지에 의한 퍼지 추론

확신도를 구하기 위하여 표준패턴과 시험패턴의 주파수 에너지에 대한 퍼지값의 소속도 함수 값을 구해야 하는데, 여기서 두 패턴 사이의 확신도 $S^e(i)$는 본 연구의 확신도 계산에 적합하도록 식 (8)와 같이 표현하여 ∧-∨(: max-min) 에 의해 구한다. [21]

$$S^e(i) = \vee (\mu^{ei}_{ref} \wedge \mu^{ei}_{test}) \qquad (5)$$

단, i = 1, 2, ⋯ , 29 (i : i 번째 채널)

이렇게 29개 채널의 주파수 에너지에 대한 확신도를 구한 후, 생성규칙의 전제가 어느 정도 만족하는가를 추론하기 위해서 29개의 확신도 값을 모두 더하여, 그 음성 신호의 확신도로 사용하게 된다.

$$S^eTOTAL = \sum_{i=1}^{i} S^e(i) \qquad (i = 1,2,\ldots, 29) \quad (6)$$

최종 확신도는 시험패턴의 확신도 값을 모든 표준패턴에 적용하여서 구하게 되는데, 미리 작성된 모든 표준패턴들에 대해 구한 확신도 값들을 비교하여 최대의 확신도를 구함으로 인식된 모음의 화자를 얻게 된다. 이때 n 은 표준패턴의 수이다.

$$SIM^{e}(n) = \underset{n}{MAX} \{S^{e}TOTAL(n)\} \quad (7)$$

Ⅳ. 실험및 고찰

1. 퍼지화 패턴 작성

실험을 위한 음성 데이타의 수집은 크게 두 단계로 나누어 하였다. 첫째, 모음 검출을 위한 초기 단계로서, 20세에서 40세 사이의 남성 화자 10명이 한국어의 단모음인 /아/, /에/, /이/, /오/, /우/, /어/을 각각 10번씩 발음한 데이타를 수집하였다. 이 음성 데이타로 부터, 각 모음을 분석하여 모음 검출 알고리듬의 초기화 임계점을 구하였다. 그리고, 각 화자의 채널 주파수을 얻기 위해서 29차 디지탈 필터를 사용하여 각 주파수의 에너지 레벤을 구하였으며, 각 화자의 모음의 변화폭을 화자인식의 기본 패턴으로 설정하였다. 그 결과, 각 화자의 모음에 대한 주파수 에너지의 변화를 구할 수 있었다.

둘째, 화자인식 실험을 위해 10개의 도시명으로 구성된 음성 데이타를 수집하여 표준패턴과 화자인식 어휘로서 사용하였다. Furui [7] 의 연구에 의하면, 시간 경과에 따라 특징량이 변하여 인식률에 영향을 미치기 때문에, 이러한 문제점을 보완하기 위해서 10 단어를 10명의 화자가 한달에 1회 5번씩 발음하였고, 5 개월에 걸쳐서 수집하였다.

음성신호 처리에 사용된 파라메터는 모음 검출을 위해서 영교차율, 대수 에너지, 정규화된 자기상관계수 등을 사용하였고, 화자인식 실험을 위해 29 채널의 주파수 에너지를 사용하였다. 여기서 표준패턴은 각 화자가 발성한 음성 신호에서 검출한 모음에서 구한 각 채널의 주파수 에너지를 퍼지값으로 표현하여 사용하였으며, 시험패턴도 같은 방법으로 추출하여 퍼지값으로 표현하여 작성하였다.

2. 화자인식 실험

화자인식 실험의 구성도를 그림 3에 나타내었다. 실험을 위해, 입력된 어휘에서 제일 먼저 검출된 모음의 주파수 에너지를 구하였으며, 각 채널 주파수 에너지를 계산하여 표준패턴으로 저장하였다. 이때 각 채널 주파수의 값은 29차 까지의 값을 사용하였다.

실험 시스템에서 A/D 변환부를 제외하고는 제시된 알고리듬을 토대로 하여 C언어와 어셈블러를 사용하여 작성하였다. 실험 결과, 입력된 어휘를 받아서 처리한 후 표준패턴과 비교하여 화자를 선별토록 하여 인식하는데, IBM PC 486 호환기종으로 약 3 - 5초의 시간이 소요되었으며, 알고리듬을 좀더 개선한다면 실시간 처리도 가능하다고 생각된다.

그림 4는 5명의 화자로부터 추출한 /아/에 대한 주파수 에너지의 차이를 보여주고 있다. 그림 4에서 보면, 주파수 영역에 따라서 개인차가 나타나기 때문에, 인식 실험의 특징량 계산시에 이 영역들에 가중치 값을 부여하여 반영하였다.

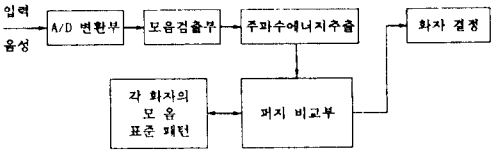

그림 3. 화자인식 시스템
Fig. 3. System of speaker recognition.

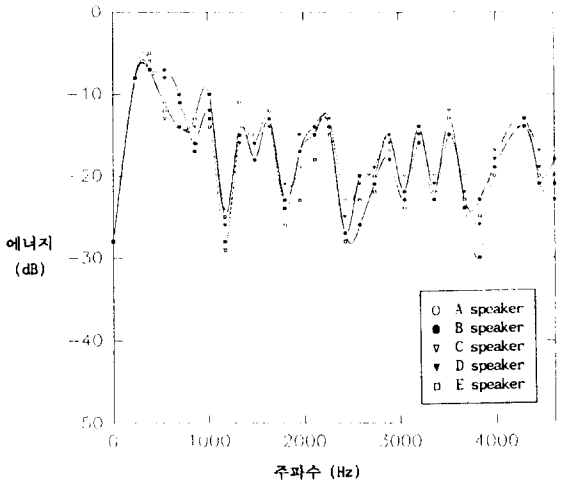

그림 4. 5명 화자의 /아/에 대한 주파수 에너지
Fig. 4. The frequency energy for /a/ vowels of 5 speakers.

3. 실험 결과 및 고찰

한국어의 음성학적 특성상 우리가 발음하는 어휘중에는 하나 이상의 모음을 가지고 있다는 점에 기초하여, 모음을 가지고 화자인식 실험을 수행 하였다. 먼저 정확한 모음을 검출하기 위해서 새로운 모음 검출 알고리듬을 제안하였으며, 모음 검출의 임계점 설

정은 대수 에너지, 영교차율, 정규화된 자기상관계수를 사용하였다.

표준패턴은 주파수 에너지를 29 프레임으로 표현하였으며, 각 프레임은 1 바이트로 구성되기 때문에 전체 29 바이트로 하나의 표준패턴을 표현하게 되며, 시험패턴의 구조도 같은 형태이다. 따라서, 한 화자에 대해 하나의 표준패턴을 구성하는 경우, 화자별로 필요한 기억장치는 29 바이트 X 10 도시명 X 5 번 = 1,450 바이트 이다. 이렇게 구성하므로, 기존에 제안된 방법들에 비해 기억용량이 대략 1/50 감소되었으며, 처리시간도 1/20 정도 단축되었다.

화자 판정을 위한 비교 방법은 기존에 사용되어온 DP, DTW 또는 HMM 방법을 탈피하여 퍼지 이론을 도입하였으며, 이로인해 음성의 특징량 변화에 대한 민감성을 대폭 감소시키는 이점등을 얻었다. 그리고 실험은 2단계로 하였는데, 먼저 텍스트 의존 화자인식 실험을 하고, 두번째로 텍스트 독립 화자인식 실험을 하였다.

표 3. "F" 화자의 인식율
Table 3. Recognition rate of "F" speaker.

화자 / 데이타	A	B	C	D	E	F	G	H	I	J	인식비	인식율 (%)
서 울						20					20/20	100
부 산						20					20/20	100
강 릉		1	1			18					18/20	90
수 원						19				1	19/20	95
인 천						20					20/20	100
제 주		,				20					20/20	100
오 산						20					20/20	100
성 남						20					20/20	100
대 구						19				1	19/20	95
안 양						20					20/20	100
평균인식율	196 / 200 \ 100 = 98 %										196/200	98

표 4. "F" 화자의 확신도
Table 4. Certainty factors of "F" speaker.

화자 / 결과		A	B	C	D	E	F	G	H	I	J
인식	수 원 J화자	0.62	0.63	0.69	0.65	0.72	0.98 •	0.66	0.62	0.61	0.67
오인식	수 원 J화자	0.61	0.62	0.67	0.66	0.70	0.95	0.68	0.64	0.63	0.97 •

• : 인식된 화자

표 3부터 표 6는 화자인식 실험에 대한 인식율들을 비교해서 보여주고 있으며, 표 3은 인식율이 가장 좋은 화자의 인식 결과이고, 표 4는 이 경우의 확신도 값의 예이다. 또한, 표 5은 인식율이 가장 낮은 경우의 예이며, 표 6에 확신도 값의 예를 나타내었다. 나머지 8명의 화자에 대한 각각의 표는 생략하고, 표 7에 종합적인 인식율을 나타내었다.

표 5. "J" 화자의 인식율
Table 5. Recognition rate of "J" speaker.

화자 / 데이타	A	B	C	D	E	F	G	H	I	J	인식비	인식율 (%)
서 울										20	20/20	100
부 산	1	1								18	18/20	90
강 릉		1		1				1		17	17/20	85
수 원										20	20/20	100
인 천	1					1				18	18/20	90
제 주		,		1				1		18	18/20	90
오 산		1		1						18	18/20	90
성 남						1		1		18	18/20	90
대 구	1					1				18	18/20	90
안 양										20	20/20	100
평균인식율	185 / 200 \ 100 = 92.5 %										185/200	92.5

표 6. "J" 화자의 확신도
Table 6. Certainty factors of "J" speaker.

화자 / 결과		A	B	C	D	E	F	G	H	I	J
인식	부 산 J화자	0.93	0.86	0.68	0.83	0.61	0.78	0.66	0.82	0.81	0.97 •
오인식	부 산 A화자	0.96 •	0.87	0.67	0.86	0.63	0.79	0.65	0.84	0.79	0.95

• : 인식된 화자

실험에 사용한 음성 데이타는 인간의 음성이 시간이 지남에 따라 특징량이 조금씩 변한다는 실험 결과에 따라, 5 개월에 걸쳐 수집하여 사용하였는데, 표준패턴을 위한 데이타는 무작위로 추출하지 않고, 1 개월에 하나씩 데이타를 수집하여 첨가하였으며, 이때 처음 얻은 음성 데이타를 표준패턴으로 한 경우, 시간이 지남에 따라 오차율이 증가하는 것을 알 수 있었다.

표 7. 텍스트 의존 화자인식 결과
Table 7. The result of text-dependent speaker recognition.

화 자	인 식 율	화 자	인 식 율
A	97.0	F	98.0
B	95.0	G	97.0
C	94.0	H	95.5
D	94.5	I	94.0
E	97.5	J	92.5

본 연구에서 제안한, 모음을 검출해서 화자를 인식하는 방법이 텍스트 독립 화자인식에 적용이 가능한 것을 증명하기 위해서, 다음과 같은 조건으로 인식 실험을 하였다. 먼저 시험패턴은 표준패턴 작성에 사용되지 않은 단어와 지하철 역 이름들을 선택하여 작성하였으며, 실험을 위해 사용한 데이타는 표 8과 같다. 실험용 데이타는 5명의 화자가 한달에 4 개씩 5 개월 동안 작성하였으며, 실험을 위해서 별도의 표준패턴을 작성하지 않았다. 실험 결과, 평균 인식율이 94.2 %을 나타 내었다. 표 9은 표 8의 데이타를 이용한 텍스트 독립 화자인식의 실험 결과를 분석한 것이다. 분석을 위해, 실험시 발생한 오인식 데이타를 화자별로 녹음한 시점을 기준으로 분류하고, 이 자료를 중심으로 시간이 경과하는데 따른 오인식 분포를 나타내었다.

표　8. 텍스트 독립 화자인식을 위한 실험 데이타
Table 8. Test data for the text-independent speaker recognition.

전 자	통 신	인 식	퍼 지	추 론
부 평	아 현	시 청	건 국	노 원

표　9. 시간 변화와 오인식율 사이의 관계
Table 9. Relationship between error rate and time variation.

오인식결과 / 화자	오 인 식 데 이 다 수					합 계
	1 개월	2 개월	3 개월	4 개월	5 개월	
A	0	1	2	4	6	13 개
B	0	1	2	5	7	15 개
C	1	1	2	3	7	14 개
D	0	0	1	3	5	9 개
E	0	0	1	2	4	7 개
오인식 합계	1	3	8	17	29	58 개
오인식율(%)	0.5 %	1.5 %	4.0 %	8.5 %	14.5 %	

V. 결 론

지금까지 화자인식 연구는 주로 텍스트 의존 방법으로 연구되어 왔다. 본 논문도 텍스트 의존 방법에서 부터 출발하였으나, 연구과정에서 발생된 결과와 문제점을 개선하여서 텍스트 독립 연구를 진행하였다. 텍스트 독립 화자인식을 위해서 미리 등록한 단어가 아닌 경우에도 화자를 인식할 수 있도록 특징 파라메터를 개발하고, 실용화가 가능하도록 처리 방법을 간략화 하였다. 이를 위해서, 화자가 발성한 음성에서 모음을 검출하여 화자인식에 사용하는 방법을 제안하였으며, 인식은 각 화자가 발성한 음성 신호에서 모음을 검출한 다음, 검출된 모음의 29 채널의 주파수 에너지를 퍼지값으로 표현한 후, 퍼지 추론을 적용하여 수행하였다.

실험을 위해 모음 검출 알고리듬을 개발하였으며, 이 알고리듬을 이용하여 입력 데이타로 부터 검출한 모음만을 인식에 사용함으로써, 실험 데이타의 크기를 1/9 정도 줄였으며, 미리 기억된 단어에 의존하지 않고 화자인식을 할 수 있었다. 실험 결과, 미리 작성된 표준패턴과 동일한 단어를 사용한 텍스트 의존 화자인식 실험은 95.5 % 인식율을 보였고, 표준패턴과 다른 종류의 단어를 사용한 텍스트 독립 화자인식 실험은 94.2 % 인식율을 보이고 있다. 따라서, 텍스트 의존 및 독립 화자인식 시스템에 모두 적용이 가능함을 알았다.

화자인식의 특징 파라메터로 29 채널 주파수 에너지를 제안하였으며, 이것은 검출된 모음으로 부터 구하였다. 실험 결과, 이 파라메터는 프레임 길이가 짧은 경우 오인식이 자주 발생하였으나, 별도의 코드북 없이 사용이 가능하고, 기존의 파라메터에 비해 인식율이 높으면서도 구성 및 계산이 간단한 특징이 있다.

앞으로 연구과제는 모음 검출을 좀 더 정확히 할 수 있도록 알고리듬을 개선하고, 피치(Pitch)를 사용해서 화자의 개인성 정보를 나타내는 파라메터를 보완하며, 신경회로망 이론의 학습 능력등을 도입하여 인식률을 높이는데 있다. 또한 연구 결과의 파라메터를 적용하는 과정을 개선한다면, 실시간 화자인식 시스템의 구현 및 각종 제품에 응용 될 수 있을 것으로 기대된다.

參 考 文 獻

[1] A. E. Roseberg, "Automatic Speaker Verification: A Review," proc.IEEE, vol. 64, no.4, pp.475-487, 1976.

[2] Atal, B., "Automatic recognition of speakers from their voices," Proc. IEEE, vol. 64, no.4, pp.460-475, 1976.

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

[4] S. Pruzansky, "Pattern-matching Procedure for automatic talker recognition," J. Acoust. Soc. Am., vol.35, 1963.

[5] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust.Soc. Amer., vol.55, no.6, pp. 1304-1312, June, 1989.

[6] M.R. Sambur, "Speaker Recognition using Orthogonal Linear Prediction," IEEE Trans. Acoust.,Speech,Signal Processing, vol.ASSP-24, pp.403-409, 1976.

[7] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," Electron. Commun. Jap., vol.57-A, pp.43-42, 1974.

[8] J.D. Markel, B.T. Oshika, A.H. Gray, JR., "Long-Term Feature Averaging for Speaker Recognition," IEEE Trans. Acoust., Speech, Signal Proc., vol. ASSP-25, pp.330-337, Aug., 1977.

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans.Acoust.,Speech,Signal Process., vol. ASSP-29, pp.254 -272, 1981.

[10] S. Furui, "Comparison of Speaker Recognition Methods using Statistical Features and Dynamic Features," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-29, pp.342-350, 1981.

[11] R. E. Helms, "Speaker recognition using linear prediction vector codebooks", Ph.D. thesis, Southern Methodist University, 1981.

[12] J.T. Buck, D.K. Burton, and J.E. Shore, "Text-dependent speaker recognition using vector quantization," in Proc. ICASSP, vol. 1, pp. 391 -394, 1985.

[13] F.K. Soong, A.E. Rosenberg, "On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition," IEEE Trans.Acoust., Speech, Signal Process., vol. ASSP-36, no.6, pp.871-879, 1988.

[14] D.K. Burton, "Text-Dependent Speaker Verification using Vector Quantization Source Coding," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-35, pp.133-143, Feb., 1987.

[15] B.H. Juang, F.K. Soong, "Speaker recognition based on source coding approaches," in Proc. IEEE Int. Conf. Trans. Acoust., Speech, Signal Processing, pp.613-616, 1990.

[16] M.S. Chen, P.H. Lin and H.C. Wang, "Speaker Identification Based on a Matrix Quantization Method," IEEE Trans. on Signal Process., vol. 41, no. 1, 1993.

[17] 中田和男, 音聲情報處理の基礎, オーム社, 1981

[18] J. D. Markel, and A. H. GRAY Jo., *Linerar Prediction of Speech*, Springer Verlang, N.Y., 1976

[19] 鈴木久喜, 音聲の 線形豫測, コロナ社, 1978

[20] 寺野 壽郎, 淺居 喜代治, 菅野 道夫, *Fuzzy System Theory and its Appl- ication*, オ- ム社, 1989.

[21] Abraham Kandel, *Fuzzy Techniques in Pattern Recognition*, John Wiley & Sons, 1982.

著者紹介

金 에 녹(正會員)

1955年 6月 15日生. 1977年 2月 광운대학교 전자계산학과 졸업(이학사). 1982年 2月 건국대학교 대학원 전자공학과 졸업(공학석사). 1993年 8月 건국대학교 대학원 전자공학과 졸업(공학박사). 1983年 12月 ~ 현재 연암공업전문대학 전자계산과 부교수. 주관심 분야는 음성인식, 화자인식, 문자인식, Fuzzy 시스템 등임.

金 炯 來(正會員)

연세대학교 이공대학 전기공학과(전자공학 전공) 및 동대학원(공학박사). 건국대학교 공과대학 학장보 및 교학과장. 美國 North western 大學校 교환교수. 日本 橫 國立大學校 객원교수. 대한전자공학회 회로 및 시스템 연구회 위원장. 대한전자공학회 학술담당 협동이사. 현재 건국대학교 공과대학 전자공학과 교수. 대한전자공학회 총무이사. 주관심 분야는 Pattern Recognition 및 Speech Recogniton.

卜 赫 圭(正會員)

1963年 9月 10日生. 1986年 2月 청주대학교 전자공학과 공학사 학위 취득. 1993年 2月 건국대학교 전자공학과 공학석사 학위 취득. 1994年 ~ 현재 건국대학교 전자공학과 박사과정 재학중. 주관심 분야는 Pattern Recognition, 화상처리, 문자인식 등임.