PVPF방법과 퍼지 이론을 이용한 한국어, 영어 및 일본어 화자 인식에 관한 연구

김 연 숙†

요 약

본 논문에서는 피치 파라미터와 퍼지 추론을 포함한 화자 인식 알고리즘을 제안한다. 시간영역에서 검출 알고리즘의 장점인 잡음에 강인함을 가진 PVPF법을 제안하여 피치를 검출한다. 또한 화자 인식에서 특징량들의 애매성을 표현하고 인식하는 방법으로 퍼지 이론을 도입하였다. PVPF는 음의 시간적인 특징을 이용하여 국부적으로 봉우리와 골을 이룬다는 것을 이용한 계산량이 적고 잡음에 강인한 피치 검출법이다.

A Study on Korean, English and Japanese Speaker Recognitions using the Peak and Valley Pitch Detection and the Fuzzy Theory

Yeoun-Sook Kim†

ABSTRACT

This paper proposes speaker recognition algorithm which includes both the pitch parameter and the fuzzy inference. This study proposes a pitch detection method PVPF(peak and valley pitch detection function) by means of comparing spectra which utilizes the transform characteristics between time and frequency.

In this paper, makes reference pattern using membership function and performs vocal tract recognition of common character using fuzzy pattern matching in order to include time variation width for non-linear utterance time.

1. 서 론

인간과 기계와의 정보 교환에 대한 필요성이 더욱 절실한 현대 산업 사회에서 인간과 기계 상호간의 통신에 관한 여러 가지 연구와 실험이 행하여져 왔다.

인간에게는 의사 전달을 위해서 음성, 기호, 제스츄어 등의 여러 가지 수단이 있지만, 음성을 이용하는 것이 가장 간편하면서도 빠르고 정확하며, 또한 자연스럽고 효율적인 방법이다. 특히, 컴퓨터와 인간의 상호 대화를 위해서 음성의 정보 전달에 관한 관심이 고조되고 있는 실정이다.[1][2]

음성 인식과 화자 인식은 인간에게서 기계로의 의사 전달 방법이다.[3] 화자 인식 연구는 1963년 Bell 연구소의 Pruzansky가 시간 평균 스펙트럼을 사용해서 화자 식별 실험을 하였고, 1970년 일본의 Furui, Saito, Itakura 등이 PARCO계수와 피치의 통계적 파라미터를 사용하여 화자 식별 실험을 하였다. 1974년 Bell 연구소의 Atal은 화자 식별 및 화자 확인 실험을 하였으며, 1981년 Bell 연구소의 Furui는 텍스트 의존 화자 인식 실험을 하였다. 국내에서는 1989년 이혁재가 결정 함수 개념을 도입하여 거리 개념에 적용시켜 화자 인식 실험을 하였으며, 1991년 권석규는 이혁재의 연

† 정 회 원 : 건국대학교 대학원 전자공학과
논문접수 : 1998년 3월 2일, 심사완료 : 1998년 12월 1일

구 결과에 DSP 칩을 사용하여 H/W 설계를 하였다. 음성 인식에 사용되는 파라미터들로는 피치 변화율(pitch contour), 에너지 변화율(energy contour), LPC 파라미터, 포만트 정보, 스펙트럼의 상관관계 등을 보통 사용하고 있다.[4][5][6][7][8] 그러나 지금까지의 연구 결과는 검색, 수사 보조 등 특별한 응용에만 주로 적용되었기 때문에 대외적인 발표를 꺼리는 원인도 있었지만, 배경 잡음이 없는 이상적인 음성신호의 경우와 대상 발음이 한정된 경우에만 만족할 만한 결과가 얻어지고 있는 실정이었다.[9] 또한 음성 인식에 사용하는 발음은 단어의 수에 제한을 받지 않아야 실용성과 일반성을 가질 수 있다.[10] 이러한 문제는 화자 인식에 사용되는 파라미터들을 통계적으로 추출, 적용함으로서 해결할 수 있다. 따라서 배경 잡음에 둔감하고 화자의 개성을 통계적으로 잘 대변해 줄 수 있는 새로운 파라미터의 제안은 음성 인식 분야에서 반드시 해결되어야 할 중요하고 필수적인 과제이다.

본 논문에서는 시간 영역에서 검출 알고리즘의 장점인 분해력을 높이고 주파수 영역에서의 장점인 잡음에 강인함을 가진 PVPF(peak and valley pitch detection function)법을 제안하여 피치를 검출한다. PVPF는 음의 시간적인 특징을 이용하여 국부적으로 봉우리와 골을 이룬다는 것을 이용한 계산량이 적고 잡음에 강인한 피치 검출법이다. 음성의 패턴 인식에서 인식 성능을 저하시키는 문제점으로는 음성의 시간 변동과 주파수 변동 등이 있다. 이러한 문제를 보완하는 방법으로 DTW방법과 멀티-템플릿 방법 등이 개발되었으나, 모두가 대량의 기억 용량과 계산량을 필요로 하는 단점이 있다. 음성의 음향적 성질은 동일한 단어라도 지닌 의미와 발음 속도가 다를 수 있으며 음성 기관에 따라 복잡하게 변하므로 동일인이라 하더라도 특징량이 다르게 나타날 수 있다. 따라서 화자 인식에 있어서 특징으로 추출되는 특징량들이 화자의 특징을 나타내는 절대적인 것으로 볼 수 없으므로 그 특징량들의 애매성을 표시해 줄 필요가 있다. 여기서 그 애매성을 표현하고 인식하는 방법으로 퍼지 이론을 도입하였다.[10]

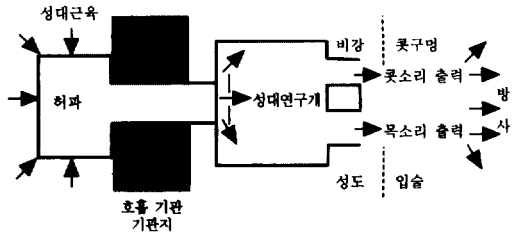
본 논문은 다음과 같이 구성되어 있다. 제 2절에서는 음성학적 분석과 음성에 대한 전 처리 과정에 대해 살펴본다. 제 3절에서는 유/무성음 검출과 국부 봉우리와 골에 의한 피치 검출법에 대해 기술하고, 제 4절에서는 피치 검출과 퍼지 이론을 이용한 화자 인식을 기술한다. 제 5절과 제 6절에서는 제안한 내용과 실험을 평가하고 결론을 맺는다.

2. 음성학적 분석과 전 처리 과정

여기서는 음성 생성원에 대한 음성학적 분석과 음성신호를 표현하는 방법에 대해 기술한다.

2.1 음성학적 분석

(그림 1)은 음성 발생 시스템의 구조적 형태를 나타낸 것으로 허파, 기관지와 호흡 기관으로 구성된 하부-성문 시스템을 포함한다. 이 하부-성문 시스템은 음성을 생성하는 에너지원이다. 인두강과 구강을 합쳐 성도(vocal tract)라 하고 비강을 비도(nasal tract)라 한다. 그러므로 성도는 후두의 출구에서 시작해서 입술에서 끝난다. 비강은 연구개로부터 콧구멍까지이다. 성도를 음성 발생 시스템의 전체를 일컫는 뜻으로는 사용하지 않으나 논의되는 특정 음에 따라서 성도와 필요에 따라서는 비강을 합한 뜻으로 사용하기도 한다. 음성 발생에 긴요한 해부학적 구성 요소를 세밀하게 나열하면 성대(vocal cord)혹은 성대 근육(vocal fold), 연구개, 혀, 이, 입술 등이다.



(그림 1) 음성 생성 시스템의 모형도
(Fig. 1) Schematized diagram of the vocal system

음성 발생 시스템이 순간적으로 변할 수 없는 것은 조음기관이 각 음을 발생할 때 유한하게 움직이기 때문이다. 이러한 조음기관은 원하는 음들을 만들어 내기 위해 그 위치를 옮기는, 사람의 조직과 근육이다. 음향 파형으로부터 얻은 주파수 도면에서도 상당한 정보를 얻을 수 있다. 모음의 경우에 강화("공진")되는 영역과 약화("반공진")되는 영역들이 스펙트럼 상에 존재한다. 이러한 공진들은 조음기관들이 다양한 음향강과 부강(subcavity)들을 성도강 내에 만들었기

때문으로, 마치 다른 길이의 오르간 파이프를 다양한 순서로 연결하는 것과 유사하다. 그러므로 주파수 도면에서 이러한 공진의 위치는 성도의 모양과 물리적 크기에 따라 결정된다. 역으로 각각의 성도 모양은 공진 주파수들의 집합으로 특정지어진다. 시스템 모델링의 관점에서 볼 때 조음기관들이 음성 시스템 필터의 성질을 결정한다. 이러한 공진이 전체 스펙트럼을 모양(form)지우므로 음성 학자들은 이를 포만트라 부른다.

원칙적으로 주어진 음에는 무한개의 포만트가 있으나, 실제로 샘플링(sampling) 후에 Nyquist 대역(일반적으로 10㎑로 샘플링하여 5㎑로까지 나타냄)에서 3~5개를 발견할 수 있다.

2.2 전 처리 과정

전형적인 음성신호의 특징은 시간에 따라 변한다는 것이다. 예를 들어, 여기(excitation)는 유성음과 무성음 사이에서 변하며, 신호의 봉우리 진폭에도 중요한 변화가 있고 음성 구간 내에서 기본 주파수도 상당히 변한다. 이는 간단한 시간 영역 처리 기법이 강도, 여기, 피치와 같은 신호의 특징과 포만트 주파수와 같은 성도 변수에 대한 유용한 표현법을 제공할 수 있음을 말해 준다. 대부분의 음성 처리 구조에서 중요한 가정은 신호의 특성이 시간에 따라 상대적으로 느리게 변한다고 하는 것이다. 이 가정은 음성신호의 짧은 구간을 분리해서 마치 정체된 특징을 갖는 연속적인 소리의 짧은 구간으로 처리하는 여러 가지 단시간(short time) 처리 기법을 사용한다. 이것은 필요에 따라 반복된다. 종종 분석 프레임이라고 불리는 단구간(short segment)은 서로 겹치게 하여 처리한다. 각 프레임에 대한 처리 결과는 하나 또는 여러 개로 나타난다. 그러므로 이런 처리는 신호 처리의 표현으로 제공할 수 있는 새로운 시간 의존성 시컨스를 생성한다.

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m) \qquad (1)$$

식 (1)에서 음성신호(필요한 주파수 대역을 추출하기 위해 선형 필터를 통과시킨 후)는 선형 또는 비선형이 되는 전달 함수 T[ ]를 필요로 하고 어떤 조정 가능한 변수 또는 변수 쌍에 의존하게 된다. 최종 시컨스는 샘플 인덱스 n에 해당하는 시간에 위치한 창함수 시컨스와 곱해진다. 영이 아닌 모든 곱셈 값을 합한다. 보통 창함수의 길이는 유한하다. Q_n값은 시컨스 T[x(m)]의 국부적으로 가중치가 적용된 평균값 시컨스이다.

단시간 에너지에 대한 간단한 정의는 식 (2)에서 나타내고 있다.

$$E = \sum_{m=n-N+1}^{n} x^2(m) \qquad (2)$$

즉, 샘플 n에서의 단시간 에너지는 n-N+1에서 n까지 N샘플의 자승 합이다. 일반적인 표현식인 식 (1)에 의해 T[ ]는 단순히 자승이 되고 창함수 w(n)은 식 (3)과 같다.

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

식 (1)의 일반적인 표현식에 대한 중요한 특징은 창함수 w(n)과 시컨스 T[x(n)]의 이산 컨벌루션 형태로 볼 수 있다. 따라서 Q_n은 임펄스 응답이 $h(n) = w^2(n)$인 선형 시불변 시스템의 출력으로 해석할 수 있다. [11][12]

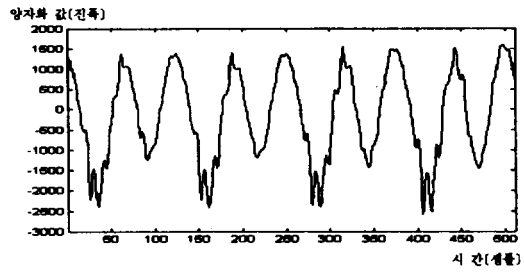
3. 유/무성음 검출과 피처 검출

여기서는 음성신호의 확률적인 분포를 사용하여 정확한 유/무성음을 검출하는 알고리즘에 대해 기술하고, 시간 영역에서 분해력을 높이고 주파수 영역에서 잡음에 강인한 피치 검출법에 대해 기술한다.

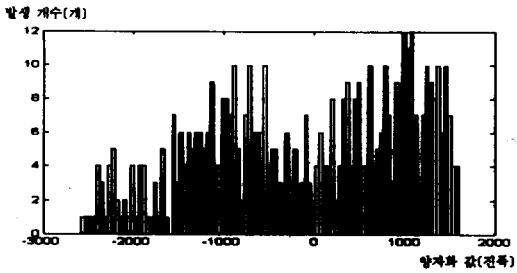
3.1 유/무성음 검출

음성신호에 있어서 유·무성음을 정확히 검출하기란 매우 어렵다. 음성신호는 일반적으로 감마 분포를 이루므로 입력 음성을 필터의 통과없이 분포도의 구성으로 하여 유·무성음 및 묵음을 검출한다.

(그림 2)는 유성음 구간에서의 히스토그램을 나타낸 것으로 전반적으로 분포가 일률적인 것이 특징이다.



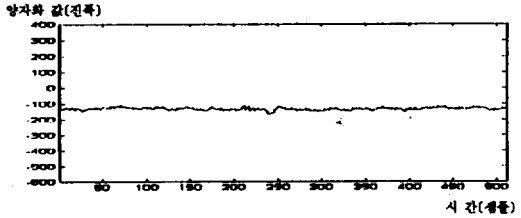
a) 음성신호의 유성음 구간



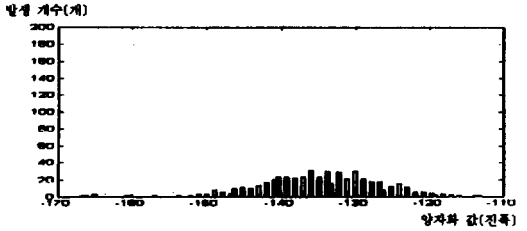
b) 유성음 구간의 히스토그램에 대한 예

(그림 2) 음성신호의 유성음 구간의 검출 예
(Fig. 2) Example of detection for voice sound edge of speech signal

(그림 3)에서 무성음 구간은 히스토그램이 평탄한 반면에 발생 수가 적고 영 부분에 몰려있다는 것이 특징이다.



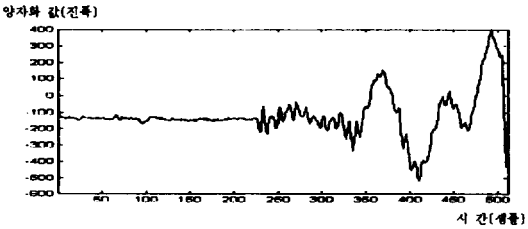
a) 음성신호의 무성음 구간



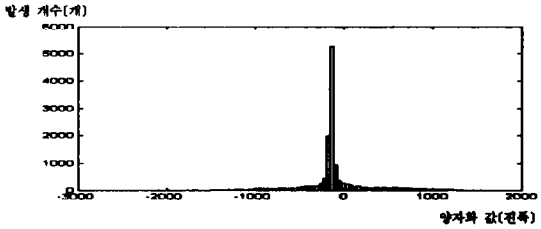
b) 무성음 구간의 히스토그램에 대한 예

(그림 3) 음성신호의 무성음 구간의 검출 예
(Fig. 3) Example of detection for unvoice sound edge of speech signal

(그림 4)에서 전이 구간은 분포가 영부분에 밀집되어 영에서 발생 수가 제일 많다.



a) 무성음에서 유성음으로 넘어가는 전이 구간



b) 무성음에서 유성음으로 넘어가는 전이 구간에 대한 히스토그램의 검출 예

(그림 4) 전이 구간에 대한 검출 예
(Fig. 4) Example of detection for excitation edge

3.2 피치 검출

일반적으로 피치 검출법은 영역별로 시간영역법, 주파수 영역법으로 나누어진다. 시간 영역법으로는 자기상관관계법이 있으며 신호의 상관관계에 따른 봉우리와 골을 강조하여 검출하기 때문에 분해력이 높으나 잡음에는 약하다. 주파수 영역법으로는 FFT를 수행하여 검출하는게 일반화되어 있다. 이 방법은 영역별 전이에 따른 계산시간이 방대하며 잡음에는 강인하다. 하지만 음의 전이 구간에서는 검출하기가 어려운 것이 단점이다. 따라서 본 논문에서는 음의 시간적인 특징을 이용하여 국부적으로 봉우리와 골을 이룬다는 것을 이용하여 계산량이 적고 잡음에 강인한 피치 검출법을 제안하였다.

음성신호를 발생원에 따라 분석을 해보면 화자의 개성을 담고 있는 기본 주파수와 성도의 필터 링 과정에서 발생되는 포만트들로 이루어져 있다. 그리고 신호에 있어서 기본 주파수의 n배 되는 고조파들의 영향은 음성신호에 있어서 봉우리와 골을 검출함으로써 그 영향을 제거할 수 있다.

봉우리와 골을 검출하는 것이 식 (4)와 같다.

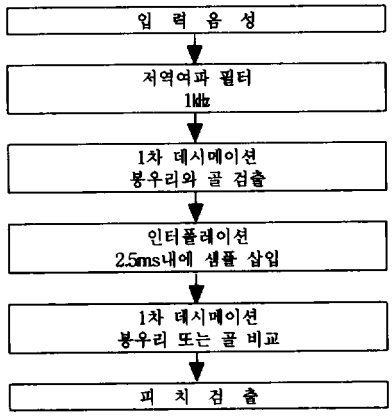
$$PV(n) = [s(n+1)-s(n)]*[s(n+2)-s(n+1)], \quad n=1,2,3,.....,k \tag{4}$$

여기서 PV(n)은 검출된 봉우리와 골들이고, S(n)은 음성신호이다. 만약 PV(n)의 값이 음의 값이면 봉우리와 골로 간주하고, 양의 값이나 영일 때는 상승이나 하강 중인 샘플로 간주한다. 이러한 간단한 식을 적용하는데 있어서 신호를 확대하여 분석을 하여 보면, 봉우리와 골로 이루어져 있다는 것을 알 수가 있다. 따라서 영이 아닌 확대된 신호에 대해서 식 (4)의 과정을 수정하여 적용하면

$$\overline{PV2} = [\ PV2 - PV1\ ] * [\ PV3 - PV2\ ]$$
$$= \begin{cases} 0, & [PV2-PV1]*[PV3-PV2] > 0 \\ 1, & [PV2-PV1]*[PV3-PV2] < 0 \end{cases} \quad (5)$$

여기서 PV1은 영이 아닌 첫 번째 PV(n)의 값이고, PV2는 영이 아닌 두 번째 PV(n)의 값이며, PV3는 영이 아닌 세 번째 PV(n)의 값이다. 그리고 $\overline{PV2}$는 검출된 봉우리와 골 값이다. 그러나 식 (5)를 여성이나 어린이의 음성에 적용할 때는 음의 형태 자체가 1계 데시메이션만으로도 검출이 가능하기 때문에 한 번의 인터폴레이션을 적용하여 데시메이션을 적용하게 된다. 인터폴레이션의 지연 값은 검출된 봉우리와 골의 2.5ms이고, 크기는 두 이웃 검출 값의 중간 값을 적용한다. 따라서 모든 신호 즉 여성, 어린이, 남성에 모두 적용할 수 있는 강인한 알고리즘을 음성신호에 적용하게 된다.

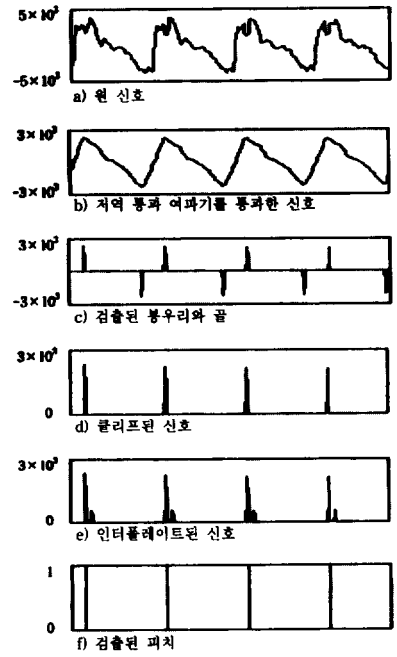
제안한 피치 검출법 PVPF(peak and valley pitch detection function)의 구성도를 (그림 5)에 나타내었다.



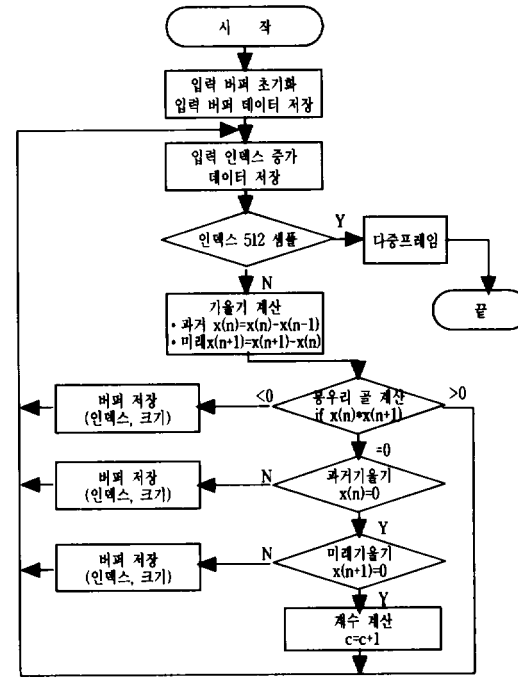
(그림 5) 제안한 피치 검출법의 구성도
(Fig. 5) Block diagram of proposed pitch detection method

PVPF의 결과는 (그림 6)에서 처럼 얻을 수 있다. 그림 a)는 한국어 모음의 한 프레임이고 b)는 저역 통과 여파기를 통과한 결과이며, c)는 1차 봉우리와 골을 검출한 결과이고 d)는 2차 데시메이션을 통해서 걸러진 봉우리와 골에 대한 결과이다. 그리고, e)는 1차 인터폴레이션을 수행한 것이고 f)는 이것을 다시 결정 논리를 통해서 피치를 결정한 결과이다.

(그림 7)은 PVPF를 흐름도로 나타낸 것이다. 입력 버퍼에 512 샘플이 채워지면 한 프레임으로 인정하여 1kHz에 해당하는 LPF를 수행한 다음에 국부 봉우리와 골을 이용한 피치 검출을 수행하게 되는 블럭도이다.



(그림 6) PVPF 알고리즘의 처리 과정
(Fig. 6) Disposal processing of PVPF algorithm



(그림 7) PVPF의 흐름도
(Fig. 7) Flowchart of PVPF

4. 피치 검출과 퍼지 이론을 이용한 화자 인식

음성신호의 피치 검출을 위해서는 퍼지 추론의 생성 규칙을 이용하여 특징 파라미터에 대한 소속도 함수를 구하고 퍼지 집합을 생성시켜야 한다.

IF 음성 신호내 모음을
  선형 필터된 후의 예비 피치 성분의 주파수 에너지 E_i가 퍼지 값 X_{Fi}를 갖고, 예비 피치 index P_i가 퍼지 값 X_{Pi}를 갖는다면
THEN
  음성 신호 "/아/, /에/, /오/, /우/, /이/"이다.

따라서 각 화자가 발성한 모음에 대해서 국부 봉우리와 골을 이용한 피치 검출법에서 얻은 피치 주파수를 사용하여 퍼지 집합을 형성한다. 이것을 예비 피치(pre-pitch)라고 정의한다. 또한 시간 영역의 신호를 주파수 신호로 변환하여 스펙트럼의 정보를 얻는다. 스펙트럼을 다시 검출된 스펙트로그램의 차를 구하여 정확한 피치의 정보를 얻는다. 여기서 얻은 정보를 사용하여 퍼지 값을 할당하고, 정규화된 주파수 에너지를 행렬 양자화를 행하여 코드 북을 생성시킨 후, 각 음성신호별로 주파수 값에 대하여 퍼지 집합의 전체 집합의 원소에 해당하는 퍼지화 값을 할당한다.

4.1 피치 검출

각 화자가 발성한 모음에 대해서 국부 봉우리와 골을 이용한 피치 검출법에서 얻은 피치 주파수를 사용하여 퍼지 집합으로 생성시킨 예비 피치에 대한 예비 피치 주파수 특징량과 정규화된 에너지 특징량의 퍼지화를 <표 1>에 나타내었다.

<표 1>의 a)는 예비 피치 존재 가능한 주파수에 대한 퍼지 값이고 b)는 예비 피치 존재 가능한 정규화된 에너지에 대한 퍼지 값으로 여기서 주파수 값을 100개의 퍼지 값으로 나타낸 것은 실제 주파수 스펙트럼 상에 존재 가능한 영역을 주파수와 개수로 나누어 사용하였고, 정규화된 에너지는 포만트 성분이 걸러진 신호에 대해서는 포락을 유지하지 않는다는 이유에서 존재 가능 개수를 계산하여 40개의 값을 할당하였다.

〈표 1〉 예비 피치 주파수 특징량과 정규화된 에너지 특징량의 퍼지화
〈Table 1〉 A fuzzified features of pre-pitch frequencies and generalized energy

a) 예비 피치 주파수 특징량

예비 피치 주파수	퍼지 값
1 - 10	0
11 - 20	2
21 - 30	3
31 - 40	4
41 - 50	5
⋮	⋮
991 - 1000	99

b) 정규화된 에너지 특징량

정규화된 예비 피치 대수 스펙트럼	퍼지 값
0	1
0.025	2
0.05	3
0.075	4
0.1	5
⋮	⋮
1	40

퍼지 이론에 의한 화자 인식은 화자가 발성한 음성에 대해 FFT를 수행한 후 행렬 양자화 인덱스와 각 주파수의 스펙트럼 양자화를 특징량으로 사용하여 음성의 변동을 해결할 수 있도록 퍼지화 패턴으로 표현한다. 따라서 스펙트럼 양자화는 주파수를 채널로 나누어 각각의 중심 주파수에 해당하는 에너지에 대해 0.1dB 마다 퍼지 값을 주어 대응시킨다.

4.2 예비 피치 에너지 특징량의 퍼지화

입력 데이터에 대해 국부 봉우리와 골이 오인식 피치일 수가 있으므로 이 자체를 예비 피치라하여 각 화자가 발성한 모음에 대해 국부 봉우리와 골을 이용한 피치 검출법에서 얻은 피치 주파수를 사용하여 퍼지 집합으로 형성시킨 예비 피치에 대한 예비 피치 주파수 특징량의 퍼지화를 <표 2>에 나타내었다. 신호에 대해 퍼지 값을 만들기 위해 주파수 1㎑를 대역 30으로 나누어 33Hz씩 분류하고 에너지는 0에서 1까지 정규화시켰다.

〈표 2〉 예비 피치 에너지 특징량의 퍼지화
〈Table 2〉 A fuzzified features of pre-pitch frequencies

퍼지 값 / 주파수[Hz]	표준 패턴		시험 패턴	
	퍼지값	에너지[dB]	퍼지값	에너지[dB]
대역 1 : 33	32	0.825	31	0.775
대역 2 : 66	30	0.750	29	0.750
대역 3 : 99	28	0.700	28	0.700
대역 4 : 132	26	0.650	23	0.575
:	:	:	:	:
대역 26 : 858	29	0.725	28	0.700
대역 27 : 891	26	0.650	26	0.650
대역 28 : 924	20	0.500	23	0.575
대역 29 : 957	16	0.400	19	0.475
대역 30 : 1000	12	0.300	11	0.275

확신도를 구하기 위하여 표준 패턴과 시험 패턴의 스펙트럼 양자화 값에 대한 퍼지 값의 소속도 함수 값을 구해야 한다. 여기서 두 패턴 사이의 확신도 $S_e(i)$는 $\wedge$-$\vee$(: max-min)에 의해 구한다.[13][14][15]

$$S_e(i) = \vee (\mu_{ei} \text{ ref} \wedge \mu_{ei} \text{ test}) \quad (6)$$

단, i = 1, 2, ... , N (i : i번째 채널)

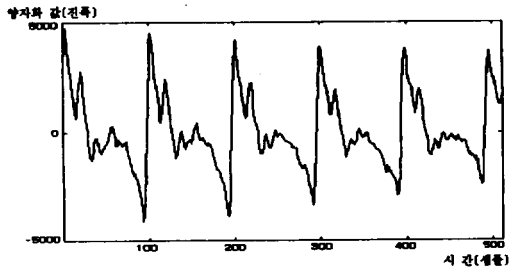
Sub-band의 스펙트럼 양자화 값의 경우 표준 패턴과 시험 패턴의 퍼지 값이 모두 같으면 확신도 $S_e(N)$은 1이 된다. 이것은 두 패턴들의 의미가 일치하는 것을 나타낸다.

<표 3>은 대역 1의 예비 피치 에너지에 대한 확신도 결과를 나타낸 것이다.

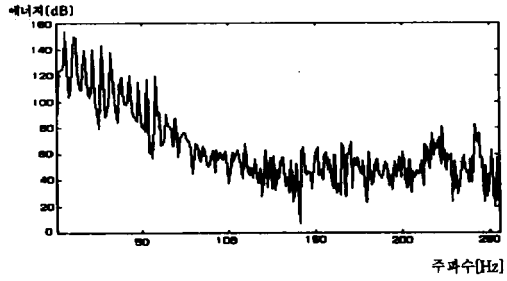
〈표 3〉 대역 1의 예비 피치 에너지에 대한 확신도 결과
〈Table 3〉 Result of certainty factors of 1st band pre-pitch energy

확신도 / 퍼지값	표준패턴의 소속도 함수	시험패턴의 소속도 함수	확신도(1)
1	0.0	0.0	0.0
2	0.0	0.0	0.0
:	:	:	
19	0.8	0.8	0.8
20	0.9	0.9	0.9
21	1.0	1.0	1.0
22	0.9	0.9	0.9
23	0.8	0.8	0.8
:	:	:	:
39	0.0	0.0	0.0
40	0.0	0.0	0.0

(그림 8)은 한국어 /아/음에 대한 한 프레임 데이터를 보인 것이다. 그리고 이것을 FFT로 주파수 성분을 나타내고 예비 피치에 대한 선형 예측 필터 처리하여 포만트 성분을 걸러내고, 피치에 대한 정보를 얻는다. (그림 9)는 한국어 /아/의 스펙트럼을 나타낸 것이다.



(그림 8) 한국어 /아/에 대한 음성신호
(Fig. 8) Speech signal for Korean /a/

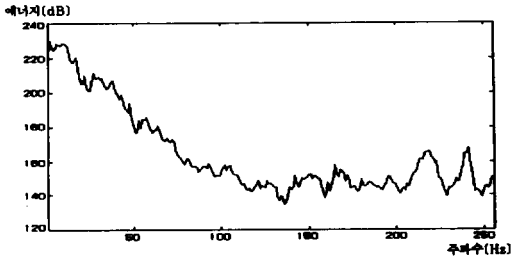


(그림 9) 한국어 /아/에 대한 스펙트럼
(Fig. 9) Spectrum of speech signal for Korean /a/

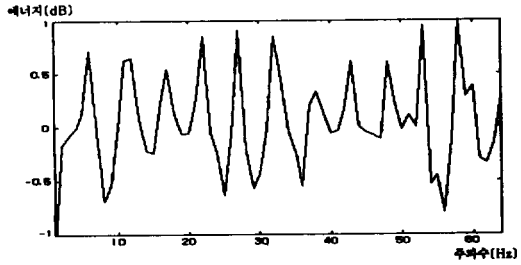
그림에서 알 수 있듯이 예비 피치를 사용하여 얻은 정보는 피치의 정확한 검출로 실제 화자에 대한 개인 정보를 검출할 수 있다.

각 화자가 10번씩 발음한 한국어 단모음 /아/, /에/, /오/, /우/, /이/를 사용해서 FFT 512 샘플을 구한다. (그림 10)은 한국어 /아/음에 대한 선형 저역 여파 필터된 피치 정보를 나타낸 것이다. (그림 11)은 예비 피치를 사용하여 선형 디지털 필터 링을 수행한 후에 원래 스펙트럼과 LPF가 수행된 신호에 대한 스펙트럼의 차 신호를 구한 값으로써 1kHz이내에 피치가 존재한다는 이론하에 검출된 피치 집합이다. 이렇게 구해진 피치 집합에 대해서 퍼지 집합을 작성하게 된다. 이것은 음성신호에 있어서 포만트 신호를 걸러내고 남은 잔여 신호에 대해서는 피치 주파수만이 존재한다는 논리하에서 처리된 것이다. 그리고 작성한 각 화자별 행렬

양자화 코드 북을 가지고, 입력된 시험 패턴에서 구한 선형 저역 여파 필터를 사용하여 코드 북의 값과 비교하여 가장 작은 오차 값을 갖는 행렬 양자화 코드 북 인덱스 번호를 구하게 된다. 이때 구한 인덱스 번호가 퍼지화 패턴 작성에 이용되며, 또한 화자 인식에서 사용된다. 이와 같이 표현되는 퍼지화 패턴을 좀 더 명확하게 나타내기 위해서, 퍼지화 패턴을 퍼지 값과 소속도 함수 값과의 관계로 설명하였다. 이때 소속도 함수 값은 퍼지화 패턴의 의미를 어느 정도 포함하고 있는가를 나타내는 것으로, 1.0에서 0.0 사이의 값으로 표현하며, 값이 1.0인 경우에는 퍼지화 패턴의 의미를 완전히 포함하는 것이며, 0.0 일 때는 완전히 포함하지 않는 것을 의미한다. 여기서 0.0과 1.0사이의 소속도 함수 값을 중심 퍼지 값에서부터 벗어나는 정도에 따라 0.05씩 감소하도록 한다. 이때 소속도 함수의 퍼지 값의 범위는 두 특징량의 변동을 흡수해 주는 정도로써, 그 범위에 따라 흡수되는 정도의 차이가 생기므로 적당한 값을 선정해 주어야 한다.

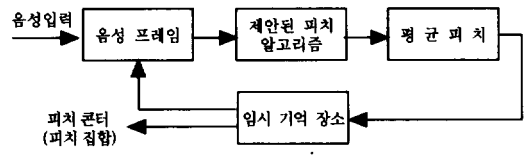


(그림 10) 한국어 /아/에 대한 예비 피치를 사용하여 저역 여파 필터를 통과한 파형
(Fig. 10) Spectrum of low pass filtered spectrum with pre-pitch for Korean /a/



(그림 11) 한국어 /아/에 대해 예비 피치를 사용한 퍼지 집합 검출 과정
(Fig. 11) Fuzzy detection using pre-pitch for Korean /a/

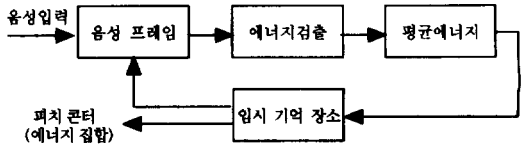
여기서 범위 설정시 퍼지 값의 범위를 너무 크게 하면 두 패턴 사이의 확신도 값이 커지므로, 어느 정도 다른 패턴도 유사성이 높다고 평가할 수 있으며, 퍼지 값이 너무 작게 설정되면 유사성이 있는 패턴도 다른 것으로 인식할 수 있으므로, 실험을 통하여 가장 높은 인식율을 보이는 퍼지 값의 범위를 구하여 사용한다.



(그림 12) 피치 콘터 검출의 블럭도
(Fig. 12) Block diagram of Pitch Contour Detection

(그림 12)는 피치 패턴을 구하는 구성도이다. 입력된 음성에 대해서 한 프레임을 512단위로 피치를 구하고 이 피치에 대해서 평균 피치를 구한다. 이렇게 음성인 구간에 대해서만 평균 피치를 얻고 전체 구간에 대해서 피치 패턴을 얻게 된다.

또한 에너지 패턴도를 구하는 것은 피치 패턴도를 구하는 것과 유사하다. 먼저 프레임별로 에너지를 구하고 평균 에너지를 구한다. 이렇게 구한 평균 에너지에서 전체 에너지 패턴을 얻는다. 이 구성도가 (그림 13)에 나타나 있다.



(그림 13) 에너지 콘터 검출의 블럭도
(Fig. 13) Block diagram of energy contour detection

퍼지 이론을 사용할 확신도를 구하기 위해서 표준 패턴과 시험 패턴의 주파수에 대한 퍼지 값의 소속도 함수를 구해야 하는데, 두 패턴 사이의 확신도 $S^c(i)$는 퍼지 추론의 합성 규칙을 적용하여 다음과 같이 표현한다.

$$S^c(i) = \vee(\mu^{ci}_{ref} \wedge \mu^{ci}_{test}) \qquad (7)$$

단, i = 1, 2, ... , n (i : 프레임 번호)

여기서,

μ^{ci} ref : i번째 프레임 코드 북 인덱스에 대한 표준 패턴의 소속도 함수

μ^{ci} test : i번째 프레임 코드 북 인덱스에 대한 시험 패턴의 소속도 함수

$S^c(i)$: i번째 프레임 코드 북 인덱스에 대한 확신도 값

N번째 프레임의 코드 북 인덱스 경우 표준 패턴과 시험 패턴의 퍼지 값이 모두 같으면 확신도 $S^c(i)$는 1.0이 된다. 이것은 두 패턴들의 의미가 일치하는 것을 의미한다. 그러나, 표준 패턴과 시험 패턴의 퍼지 값이 같지 않을 경우 프레임의 코드 북 인덱스는 최대를 취한다. 같은 방법으로 전체 프레임에 대해서 확신도를 구한 후, 생성 규칙의 전제가 어느 정도 만족하는가를 추론하기 위해서 확신도 값을 모두 더하여 그 음성신호의 확신도를 사용하게 된다.

$$S^c{}_{TOTAL} = \sum_{i=1}^{n} S^c(i) \quad (i = 1, 2, 3 \ldots, n) \qquad (8)$$

이와 같이 모음을 인식하기 위해서는 프레임별로 확신도를 계산해야 하는데, 이때 모음의 프레임 길이가 불규칙하므로 각 프레임별로 구한 확신도를 모두 누적한 후, 전체 프레임 수로 나누어줌으로써 표준 패턴에 대한 시험 패턴의 확신도를 구하게 된다.

$$SIM^c = \{ \sum_{j=1}^{j} S^c{}_{TOTAL}(j) \} / j \qquad (9)$$

(단, j : 전체 프레임 수)

이러한 과정을 1개의 시험 패턴에 대해 모든 표준 패턴에 적용하면, 표준 패턴을 구성하고 있는 각 패턴들에 대한 최종 확신도를 구할 수 있다. 여기서 최종 확신도들에 대한 최대의 확신도를 구함으로써 인식된 모음의 화자를 얻게 된다. 이때 n은 표준 패턴의 수이다.

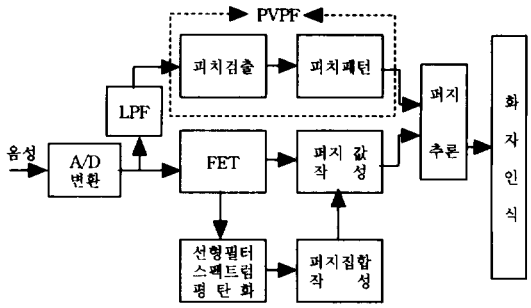
$$SIM^c(n) = \max_{n} \{SIM^c\} \qquad (10)$$

(단, n : 표준 패턴의 수)

5. 실험 및 고찰

제안된 인식 시스템은 크게 세 가지로 나누어 수행이 된다.

먼저 유성음에 대해서 전처리 과정으로 저역 여파 필터를 거친다. 이 검출된 음성에서는 프레임별로 에너지와 피치 정보를 조사하여 에너지 콘터와 예비 피치를 작성한다. 두 번째로 원래 음성과 저역 여파된 입력 음성에 대해서 퓨리에 변환을 사용하여 시간 영역에서 주파수 영역으로 변환한다. 이 두 스펙트럼 간의 차를 얻어 평탄화된 스펙트럼을 얻는다. 이렇게 구한 스펙트럼 값은 주파수와 정규화된 에너지 값으로 양자화하여 예비 피치 주파수에 해당하는 값을 코드화하여 퍼지 값을 만들고 이 퍼지 값을 사용하여 표준 퍼지 패턴을 만든다. 마지막으로 두 과정에서 작성된 예비 피치를 사용한 퍼지 패턴을 사용하여 실험하였다. 인식에 사용된 데이터는 10kHz로 샘플링되고 16bit로 양자화한 시료를 사용하였다. 화자 인식 실험의 구성도가 (그림 14)에 나타나 있다.



(그림 14) 화자 인식 실험의 구성도
(Fig. 14) Block diagram of Proposed Recognition Algorithm

본 논문에서 제안한 PVPF와 퍼지를 적용하여 인식에서 발음이 같은 한국어, 영어 및 일본어의 모음을 검출하여 각 나라(한국, 미국, 일본)의 인식율을 수행하여 화자 인식 실험을 하였다. <표 4>, <표 5>, <표 6>은 각 나라(한국, 미국, 일본)의 화자가 발음한 모음을 피치 패턴과 퍼지 이론을 이용하여 인식한 인식율을 나타낸 것이다

〈표 4〉 피치 패턴과 퍼지 이론을 이용한 한국인 화자가 모음을 발음한 인식율
〈Table 4〉 Vowels recognition rate for Korean speakers using pitch pattern and fuzzy theory

입력패턴 \ 표준패턴	아	에	오	우	이	총 합	오차율(%)
아	94					94/100	6
에		92				92/100	8
오			94			94/100	6
우				93		93/100	7
이					95	95/100	5
인 식 율(%)	93.6 %					478/500	6.4

〈표 5〉 피치 패턴과 퍼지 이론을 이용한 미국인 화자가 모음을 발음한 인식율

〈Table 5〉 Vowels recognition rate for American speakers using pitch pattern and fuzzy theory

입력패턴 \ 표준패턴	a	e	o	u	i	총 합	오차율(%)
a	95					95/100	5
e		93				93/100	7
o			96			96/100	4
u				95		95/100	5
i					93	93/100	7
인 식 율(%)	94.4 %					472/500	5.6

〈표 6〉 피치 패턴과 퍼지 이론을 이용한 일본인 화자가 모음을 발음한 인식율

〈Table 6〉 Vowels recognition rate for Japanese speakers using pitch pattern and fuzzy theory

입력패턴 \ 표준패턴	あ	え	お	う	い	총 합	오차율(%)
あ	92					92/100	8
え		90				90/100	10
お			91			91/100	9
う				94		94/100	6
い					91	91/100	9
인 식 율(%)	91.6 %					458/500	8.4

<표 7>, <표 8>, <표 9>는 LPC 켑스트럼을 사용했을 때의 기존의 인식율과 본 논문에서 제안한 피치 검출과 퍼지 추론을 사용했을 때의 인식율을 나타낸 것이다.

표에서 나타났듯이 각 나라(한국, 미국, 일본) 화자가 발음한 모음 인식율은 평균 2.13% 개선되었음을 확인할 수 있다.

〈표 7〉 기존의 방법과 제안된 방법의 한국인 화자가 발음한 모음 인식율 비교와 개선율

〈Table 7〉 Recognition rate comparison and improvement rate of vowels for Korean speakers using existing method and proposed method

시료 \ 방법		기존의 방법 (LPC Cepstrum)	제안된 방법 (피치+퍼지추론)	인식 개선율
데이터	아	92 %	94 %	2 %
	에	93 %	95 %	2 %
	오	90 %	95 %	5 %
	우	94 %	94 %	0 %
	이	90 %	96 %	6 %
평 균		91.8 %	94.8 %	3 %

〈표 8〉 기존의 방법과 제안된 방법의 미국인 화자가 발음한 모음 인식율 비교와 개선율

〈Table 8〉 Recognition rate comparison and improvement rate of vowels for American speakers using existing method and proposed method

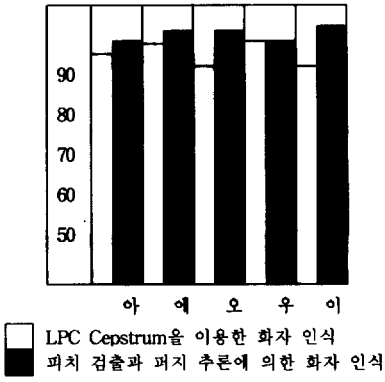
시료 \ 방법		기존의 방법 (LPC Cepstrum)	제안된 방법 (피치+퍼지추론)	인식 개선율
데이터	a	92 %	96 %	4 %
	e	92 %	96 %	4 %
	o	93 %	93 %	0 %
	u	94 %	94 %	0 %
	i	93 %	95 %	2 %
평 균		92.8 %	94.8 %	2 %

〈표 9〉 기존의 방법과 제안된 방법의 일본인 화자가 발음한 모음 인식율 비교와 개선율

〈Table 9〉 Recognition rate comparison and improvement rate of vowels for Japanese speakers using existing method and proposed method

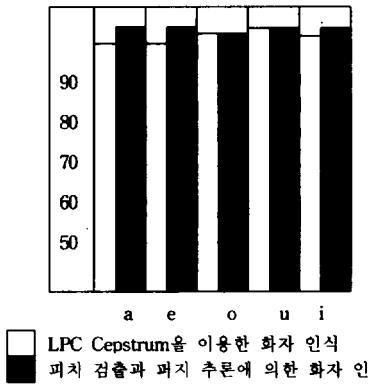
시료 \ 방법		기존의 방법 (LPC Cepstrum)	제안된 방법 (피치+퍼지추론)	인식 개선율
데이터	あ	90 %	93 %	3 %
	え	93 %	93 %	0 %
	お	90 %	90 %	0 %
	う	91 %	92 %	1 %
	い	91 %	94 %	3 %
평 균		91 %	92.4 %	1.4 %

(그림 15), (그림 16), (그림 17)은 각 나라(한국, 미국, 일본)의 모음에 대해 LPC 켑스트럼 방법, 피치 검출과 퍼지 추론 방법을 이용했을 때 각각의 화자 인식율을 막대 그래프로 나타낸 것이다.

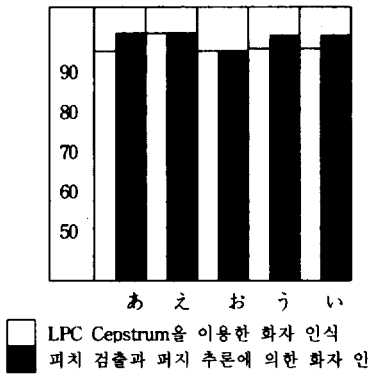


(그림 15) 인식 방법별 한국어 모음 인식율 비교

(Fig. 15) Comparison of recognition rate of Korean vowels according to methods



(그림 16) 인식 방법별 영어 모음 인식율 비교
(Fig. 16) Comparison of recognition rate of English vowels according to methods



(그림 17) 인식 방법별 일본어 모음 인식율 비교
(Fig. 17) Comparison of recognition rate of Japanese vowels according to methods

6. 결 론

본 논문의 특징은 제 3절과 제 4절에서 유/무성음 검출과 국부 봉우리와 골에 의한 피치 검출법으로 피치 검출과 퍼지 이론을 이용한 화자 인식을 제안함으로써 기존의 인식보다 우수한 결과를 얻을 수 있음을 확인한 것이다.

음성으로부터 특징을 추출하기 위해 입력 음성에 대해 시간 영역에서 국부 봉우리와 골을 사용하여 예비 피치를 구한다. 여기서 얻은 예비 피치는 퍼지 이론에 적용하기 위해서 버퍼에 저장한다. 다음으로 음성 파형은 시간 축에 기준하여 많은 변화량을 갖고 있으며 데이터 양도 많으므로 주파수 영역으로 변환시켜 특징을 추출하기 위해 퓨리에 변환을 이용하였다. 퓨리에 변환은 시간 축에서 안정된 신호를 분석하는데 중요하지만 실제로 이러한 성질을 만족하지 못한다. 따라서 음성신호를 주파수 영역으로 변환시킬 때에는 안정된 특성을 어느 정도 만족할 수 있는 구간(예를 들면 10~30ms) 단위로 분석한다. 이와 같이 주파수 영역의 특징을 추출하기 위해 FFT를 사용한다. 음성이 구강(vocal tract)으로부터 발성된다는 사실을 근거로 구강의 형태를 필터로 가정하고, 그 필터 계수를 음성의 특징으로 삼는 것이다.

여기서 얻은 스펙트럼에 예비 피치로 얻은 주파수를 사용하여 디지털 주파수 필터에 적용한다. 이렇게 주파수 필터를 사용하여 얻은 스펙트럼과 원래 신호의 차를 구하여 차 신호에 대해서 퍼지 집합을 구한다.

앞으로의 연구 과제는 화자 인식 알고리즘으로 하드웨어를 구현하여 실시간 화자 인식 시스템이 실용화되도록 계속적인 연구가 이루어져야 할 것이다.

참 고 문 헌

[1] Zemlin, W., *Speech and Hearing Science, Anatomy and Physiology,* Englewood Cliffs, N. J., Prentice Hall, 1968.

[2] L. D. Erman, "An Environment and System for Machine Understanding of Connected Speech," Ph. D. Dissertation, Carnegie-Mellon Univ., Pittsburgh, PA, 1975.

[3] Liberman, P. *Intonation, Perception, and Language,* Cambridge, Mass. : MIT Press, 1967.

[4] Dunn, H. K., "Methods of measuring vowel formant bandwidths," *JASA* Vol.33, pp.1737-1746, Dec. 1961.

[5] M. R. Sambur and L. R. Rabiner, "A Speaker Independent Digit-Recognition System," *Bell Syst. Tech. J.,* Vol.54, No.1, pp.81-102, January 1975.

[6] J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* 2nd Ed., Springer Verlag, N. Y., 1972.

[7] Tetsunori Kobayashi and Hidetoshi Sekine, "Statistical Properties of Fluctuation of Pitch Intervals and its Modeling for Natural Synthetic Speech," IEEE Trans. Acoust., Speech and Signal Proc., Vol.ASSP-32, No.1, pp.321-324, Feb. 1990.

[8] Per Hedelin and Dieter Huber, "Pitch Period

Determination of Aperiodic Speech Signals," IEEE Trans. Acoust., Speech and Signal Proc., Vol.ASSP-32, No.1, pp.361-364, Feb. 1990.

[9] Jean-Claude Junqua, Jean-Paul Haton, "Robustness in Automatic Pitch Recognition," Kluwer Academic Publishers, 1996.

[10] 김연숙, "피치 정보를 이용한 격리 단어 인식에 관한 연구", 한국학술진흥재단, SEPTEMBER 1995.

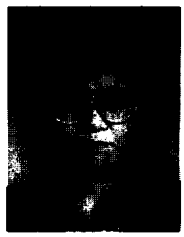
[11] J. M. Baker, "A New Time-Domain Analysis of Human Speech and Other Complex Waveform," Ph. D. Dissertation, Carnegie-Mellon Univ., Pittsburgh, PA., 1975.

[12] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Prectice-Hall, Inc., Englewood Cliffs, N. J., 1975.

[13] S. Yasunobu and S. Miyamoto, "Automatic train operation by predictive fuzzy control," in Industrial Applications of Fuzzy Logic Control, M. Sugeno, Ed., Amsterdam : North-Holland, 1985.

[14] T. Takagi and M. Sugeno, "Derivation of Fuzzy control rules from human operator's control actions," in Proc. of the IFAC Symposium on Fuzzy Information, Knowledge Representation, and Decision Analysis, Marseille, France, July 1983.

[15] J. Maiers and Y. S. Sherif, "Applications of Fuzzy set theory," IEEE Trans. Syst., man, and Cybernet., Vol.SMC-15, No.1, pp.175-189, 1985.



김 연 숙

e-mail : ysook@seoul-vos.ed.seoul.kr

1981년 아주대학교 공과대학 전자공학과 졸업(공학사)

1983년 아주대학교 일반대학원 전자공학과(공학석사)

1998년 건국대학교 일반대학원 전자공학과(공학박사)

1984년~1994년 서울동덕여고

1992년~1994년 교육부 제6차 교육과정 개정 심의위원

1992년~1996년 인천대학교 공과대학 전자계산학과 강사

1993년~1996년 서일대학 전자계산학과 강사

1992년~현재 교육부 제1종 도서편찬 집필진

1994년~현재 서울직업학교

1996년~현재 교육부 제1종 교과용 도서심의회 위원

관심분야 : 화자인식, 음성인식, 패턴인식, 퍼지추론, 신경회로망, VLSI Testing, VLSI Design Automation 등