論文
2-4-6

피치계수를 이용한 화자인식에 관한 연구
(A study on the Speaker Recognition using the Pitch)

김 에 녹*
(E-Noch Kim)

요 약

본 연구에서는 적응 공명 이론(ART2) 모델을 이용하여 화자인식 실험을 수행하였으며, 모음 검출을 통하여 미리 등록된 단어가 아닌 경우에도 화자를 인식할 수 있도록 특징 파라메터를 개발하였다. 실험을 위해 0에서 9까지의 숫자음을 남성화자와 여성화자 각각 5명씩 발음하여 사용하였으며, 이들 음성 데이터로부터 모음을 추출한 다음 얻어진 피치 계수, 선형예측 계수, 선형예측 켑스트럼 계수를 신경망의 입력 패턴으로 입력시켜 인식 성능을 측정하였다. 실험 결과 피치를 사용하는 것이 텍스트-의존, 텍스트-독립 화자인식 모두에서 다른 계수들을 사용하는 것보다 우수한 성능을 보이고 있다.

ABSTRACT

In this thesis, we perform the experiment of speaker recognition by identifying vowels in the pronunciation of each speaker using Adaptive Resource Theory 2(ART2) model. The 5 adult males and 5 adult females pronounce from 0 to 9 digits. We extract the vowels from the pronunciation of each speaker first, we are extracted characteristic coefficient through a pitch detection algorithm, a LPC analysis, and a LPC cepstral analysis to generate an input pattern of ART2. The experimental results showed that pitch coefficients are somewhat more enhanced than LPC or LPC cepstral coefficient.

1. 서 론

음성으로 화자를 자동 파악하는 화자인식 방법은 화자확인과 화자식별[1]로 나눌 수 있으며, 화자확인은 특정인이라고 자칭하는 인식 대상이 본인인지 여부를 판정하는 것이며, 화자식별은 발화자가 등록된 화자 중에서 누구인지를 판단하는 것이다. 이때 사용되는 화자 음성의 언어적 내용이 미리 등록한 언어적 내용과 동일지 여부에 따라 텍스트 의존(text-dependent) 화자인식, 텍스트 독립 (text-indep endent) 화자인식으로 구분한다[2].

현재 주로 연구되고 있는 화자인식 기법에는 입력 음성신호에서 스펙트럼 포락, 기본 주파수, 전력 등의 시간 패턴(특징 파라메터)를 추출하여 그대로 사용하는 패턴정합(Pattern Matching) 방법과 긴 시간 패턴으로부터 추출한 통계적 성질을 이용하는 방법, 그리고 표준패턴을 만들지 않고 임의의 인식 모델에 가해진 입력 음성의 특징 파라메터를 추출하여 유사군으로 분류하여 인식하는 유사군 분류법(Clustering Method) 등이 있다.

* 종신회원 : 연암공업대학 컴퓨터정보기술과 교수

논문접수 : 2001. 4. 16.
심사완료 : 2001. 4. 30.

최근에 발표된 화자인식 연구로는, 독일 Ulm 대학에서 630명의 화자가 음소가 다양하게 포함된 10개 문장을 발음한 음성 데이터베이스를 이용하여 실험한 결과 문장당 97.3% 인식 결과를 발표하였고, 미국 Rutgers 대학에서는 YOHO 데이터베이스를 사용하여 실험한 결과 남성화자는 0.09%, 여성화자의 경우는 0.31%의 오식률을 보이고 있다고 발표하였고, AT&T의 경우도 평균 오인식률 0.4%의 결과를 발표하였다. 국내의 경우 KAIST에서는 실험실 환경에서는 1.4%, 전화 음성에서는 4.8%의 오인식률 결과를 발표하였다. 현재 상용화된 화자확인 시스템들이 발표되고 있는데, 대표적인 것으로는 ITT사의 SpeakerKey(오인식률 1.7%), T-NETIX사의 SpeakEZ(오인식률 5%), Texas Instrument사의 출입구통제 시스템(오인식률 1.6~0.42%), MOSCOM사의 MOSCOM Speaker Verification (오인식률 1~2%) 등이 있다.

결과가 보고된 화자인식 시스템의 대부분은 텍스트-의존 화자인식이 주류를 이루고 있으나, 본 연구에서는 텍스트-의존뿐만 아니라 발성 내용을 고정하지 않는 텍스트-독립 화자인식 실험을 하였다. 또한 기존의 실험에서 개별적으로 사용된 바 있는 피치와 선형예측 계수, 선형예측 스펙트럼 계수를 하나의 음성 데이터로부터 각각 추출하여 신경망의 입력패턴으로 사용하였으며, 어떤 특징 파라메터를 사용하는 것이 성능이 우수한지 비교하였다.

본 연구에서 텍스트-독립 화자인식 실험을 위해 사용한 데이터들은 0에서 9까지의 숫자음으로, 음성 데이타를 내용에 구분 없이 프레임 단위로 분석하여 모음을 검출하였으며, 특징 파라메터를 추출하기 위해 인간의 성도를 모델링하여 음성을 분석하는 피치, 선형예측 계수, 선형예측 스펙트럼 계수들을 적응 공명 이론(ART2) 신경망 모델의 입력패턴으로 입력시켜 인식 성능을 측정하였다.

실험 결과 개인마다 거의 고유한 피치 계수를 사용하는 것이 선형예측 계수를 사용하는 것보다 인식 성능이 4%이상 우수함을 알 수 있었으며, 화자인식 실험의 결과도 피치계수를 사용하는 것이 다른 계수들을 사용하는 것보다 근소하나마 성능이 우수한 것으로 나타났다.

2. 전처리 및 음성신호 분석

실험을 위한 화자인식 시스템은 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, 자기상관계수 파라메터를 사용하였고, 각 단어에서 검출한 모음을 사용해서 주파수 대역을 29차로 나누어 각 채널의 주파수 에너지를 구하여 인식실험을 수행하였다.

실험을 위한 전처리 과정에서 5KHz의 저역 필터를 사용하여 고주파 성분을 제거한 뒤, 10KHz로 샘플링을 하였으며, A/D 변환기를 사용하여 음성을 디지털화 하였다. 또한 데이타의 불연속으로 인한 주파수 성분의 왜곡을 보상하기 위해 헤밍 윈도우를 사용하였으며, 모음이 각 화자의 성도 특성을 잘 나타내고 있기 때문에[3], 분석된 프레임 속에서 모음을 검출하여 화자인식 시스템에 적용시켰다.

2.1 모음 검출

한국어의 모음은 발음하는 음성마다 하나 이상을 포함하고 있으며, 모음은 주로 유성음으로 보통 단독으로 발음되는 경우와 초성 또는 종성에 자음을 동반하는 경우로 분류할 수 있다. 따라서, 정확한 모음을 검출한 후, 각 화자별로 모음의 특징을 찾아 실제 음성에 적용하여 모음 검출 알고리즘을 구성하였다.

음성신호 분석은 프레임 단위로 처리하기 때문에, 12.8 ms를 한 프레임으로 간주하여 6.4 ms 씩 겹쳐 음성신호를 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 0으로 재설정하여 다시 프레임 Counter를 증가시키는 방법으로 정확한 모음 검출을 시도하였다.

모음 검출에 사용된 알고리즘은 먼저 발음한 단모음을 분석하여 가상의 모음 시작점의 임계점을 설정한 뒤, 실제 음성에 적용하여 임계값을 변화시키는 방법을 택한 후 적당하다고 생각되는 임계값을 다른 화자의 음성신호에 적용하여 모음을 검출하도록 하였다.

IZCR = 1/N (ΣTZCR[i] * δ)
IENG = 1/N (ΣTENG[i] * δ)
IACR = 1/N (ΣTACR[i] * δ)
ICEP = 1/N (ΣTCEP[i] * δ)

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

여기서 N은 처음 설정된 프레임에서부터 임계조건을 만족하지 않는 프레임의 갯수를 나타내고 있고, δ는 선택된 임의의 임계치에 대한 변화율을 비율로서 나타낸 값으로 0.1 씩 증가시키는 값이다.

2.2 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 주파수로, 보통 모음에는 몇 개의 포만트 주파수가 있으며, 이 포만트는 발성자, 성별, 연령 등에 따라 큰 폭으로 변동하므로, 화자의 특징을 표현하는 파라메터의 하나로 포만트 주파수의 전력 스펙트럼을 변형시켜서 사용하였다.

포만트 측정 방법에는 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[4, 5]. 선형예측에 의한 포만트 주파수 추출법에는 root-solving법과 peak-picking법이 있는데, root-solving법의 계산이 복잡하고 시간이 많이 소요되기 때문에, 실험에서 포만트 주파수 추출은 peak-picking법을 사용하였다.

2.3 피치 분석

피치 정보의 특징은 개인성으로, 동일한 발음이라도 개인에 따라 다르게 나타난다. 따라서 이러한 피치 특성을 이용하면 화자인식이나 화자식별에서 좋은 결과를 얻을 수 있다. 피치 검출법들은 대개 음성 파형의 주기성을 강조하고 적절한 결정 논리를 적용하여 피치를 검출하게 된다. 주기성 강조는 성도의 공명 현상에 의해 나타나는 포만트들의 영향을 제거시키고 여기원의 피치만을 강조하는데 그 목적이 있기 때문에, 성도의 영향을 제거시키면 결정 논리가 가능해 진다.

피치 검출의 결정 논리법에는 강조된 음성 파형의 기본 주기를 실험적인 임계값이나 가중치를 적용하여 결정하는 것이 일반적이다. 결정 논리가 실험적인 임계값에 의존하면 검출 오차가 확률적으로 커지기 때문에 경우의 수에 따라서 결정 임계값을 신중히 고찰해야 한다[6]. 그러나 피치를 정확히 추출하는 방법은 성대의 진동이 항상 정확한 주기성을 갖지 않고, 음성 파형에서 성대 소스 신호를 성도의 영향으로부터 분리시켜 추출하기 어려우며, 피치의 변동 범위가 크다는 점 등으로 인하여 아직 확립되지 못하고 있다.

시간 영역에서의 피치 검출법으로는 1967년 레디(Reddy)에 의해서 제안된 국부 봉우리(peak)와 골(valley)을 이용한 방법이 사용되었으며, 그 이후에는 라비너(Rabbiner)에 의한 병렬 처리법이 주로 사용되어 왔다[7]. 그러나 이 방법은 많은 조건을 주어 처리하기 때문에 분석시 오차가 수반되어 실제 피치 이외에 having(half pitch error), doubling(double pitch error) 등이 발생하고 처리 과정의 복잡한 결정 논리로 인하여 많은 계산 시간이 요구된다.

피치 검출 방법을 종류별로 분류하여 보면, 파형 처리법에는 병렬처리 방법, 데이터 감소법, 영교차 계수법이 있고, 상관관계 처리법으로는 자기상관 방법, 변형된 상관 방법, 단순화된 역필터 추적법, 평균차 함수법(AMDF) 등이 있다. 또한 스펙트럼 처리법으로는 켑스트럼 방법과 주기 히스토그램 방법이 있다[8].

일반적인 피치 검출 결정 알고리즘 구현은 전처리와 후처리 과정을 거치게 되는데, 전처리기는 피치 검출을 용이하게 하기 위한 데이터 삭감 기능을 수행하며, 후처리기는 보정, 오류 검출, 얻어낸 피치 파형의 평탄화 등을 수행한다.

음성신호에는 반복적인 주기 정보 이외에 고주파 성분의 파형 정보가 포함되어 있는데, 이와 같은 고주파수 성분의 영향으로 인하여 피치 검출에 오류가

발생할 수 있다. 따라서 피치 검출에 불필요한 고주파 신호 성분을 제거하고 음성의 저주파수 성분만을 가지고 피치를 검출하는 방법이 단순화된 역필터 추적(Simplified Inverse Filter Tracking) 방법(으로, 본 연구에서는 음성 파형을 down-sampling 한 후, 스택트럼 평탄화를 위한 LPC 분석)을 사용하였다.

본 연구에서는 단순화 역필터 추적법을 사용하여 모음 부분에서 피치 추정을 통해 피치를 검출한 후, 이를 특징 벡터화하여 신경망의 또 다른 입력 패턴으로 사용하고, 성도 모델링을 통해서 추출하게 되는 LP 계열의 특징 벡터들과 교차 적용하여 성능을 비교하였다.

3. 적응 공명 이론(ART2) 모델

본 연구에서 사용한 적응 공명 이론(ART) 신경망 모델은 이전에 학습된 패턴을 유지하면서 새로운 패턴을 학습하는데 필요한 유연성을 잃지 않도록 설계되었다. 이 모델은 입력과 출력 노드만으로 구성되어 있으며, 처음 입력을 첫 번째 클러스터들에 대한 패턴으로 선택하고, 다음 입력을 인가하여 이전 클러스터와의 거리를 구한 후 이전 패턴들과 비교하여, 임계값보다 작으면 그 패턴의 클러스터가 되게 하고, 크면 새로운 클러스터를 계속 형성해 나가는 자율 학습 모델이다.

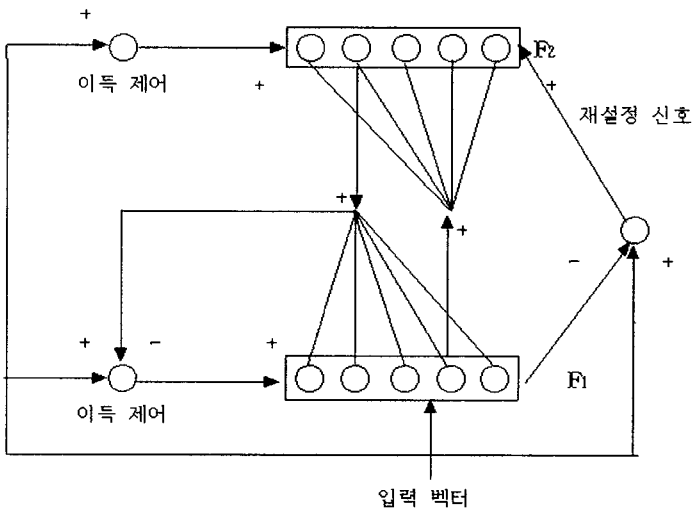
각 계층상에 있는 노드들은 다른 계층상에 있는 노드들과 완전하게 연결된다. [그림 1]의 + 표시는 활성 연결을 - 표시는 억제성 연결을 나타내며, 입력층은 출력층과 순방향 연결을 이룬다. ART 모델은 기존의 입력과 전혀 다른 입력이 인가되는 경우, 새로운 클러스터를 형성하므로 출력층의 수가 입력에 따라 무한히 커지게 된다[9].

적응 공명 이론에는 입력 패턴이 이진수인 경우에만 처리가 가능한 ART1 이진수 이외의 연속된 값도 처리할 수 있는 ART2, 화학적 전달 물질을 ART에 결합한 ART3, 그리고 지도 학습 형태를 사용하는 ARTMAP 등이 있는데, 본 연구에서는 ART2 모델을 사용한다[10, 11].

본 연구에서 사용한 ART2 모델의 처리 과정은 다음과 같다.

① 모든 계층과 부계층의 출력을 0 벡터들로 초기화하고, 1로 초기화된 사이클 카운터를 설정한다.

② 입력 패턴 I를 F_1의 w 계층에 적용한다.
이 계층의 출력은 $w_i = I_i + au_i$ 이다.



[그림 1] 적응 공명 이론 모델의 개념도

[Fig. 1] Concept of ART model

③ x 부계층을 향해 전달한다.

$$x_i = \frac{w_i}{e + \|w\|}$$

④ v 부계층을 향해 전달한다.

$$v_i = f(x_i) + bf(q_i)$$

여기서, 우변의 두 번째 항은 첫 번째 시도할 때는 0이다.

⑤ u 부계층을 향해 전달한다.

$$u_i = \frac{v_i}{e + \|v\|}$$

⑥ p 부계층을 향해 전달한다.

$$p_i = u_i + dz_{ij}$$

이때, F_2 상의 j 번째 노드는 경쟁에서 승리한 것이다. 만약, F_2가 비활성이라면, $p_i = u_i$

⑦ q 부계층을 향해 전달한다.

$$q_i = \frac{p_i}{e + \|p\|}$$

⑧ F_1 상에서의 값을 안정화시키기 위하여 ②~⑦번 과정을 반복한다.

⑨ r 계층의 출력을 계산한다.

$$r_i = \frac{u_i + cp_i}{e + \|u\| + \|cp\|}$$

⑩ 만약 $p/(e+\|r\|) > 1$이면, 리셋 기호를 F_2로 보낸다. 어떤 활성 F_2 노드를 경쟁을 위해 부적합한 것으로 표시하고, 사이클 카운터를 1로 재설정하며, 단계 ②로 돌아간다.

⑪ p 부계층의 출력을 F_2 계층으로 전달한다. F_2로의 신경망 입력을 계산한다.

$$T_j = \sum_{i=1}^{M} p_i z_{ij}$$

⑫ 승리한 F_2 노드만이 0이 아닌 출력을 갖는다.

$$g(T_j) = \begin{cases} d, & T_j = \max\{T_k\} \\ 0, & \text{otherwise} \end{cases}$$

사전에 재설정 신호에 의해 부적합한 것으로 표시된 어떤 노드도 경쟁에 참여할 수 없다.

⑬ ⑥~⑩ 단계를 반복한다.

⑭ 승리한 F_2 유니트 하의상달 방식의 가중치들을 변경한다.

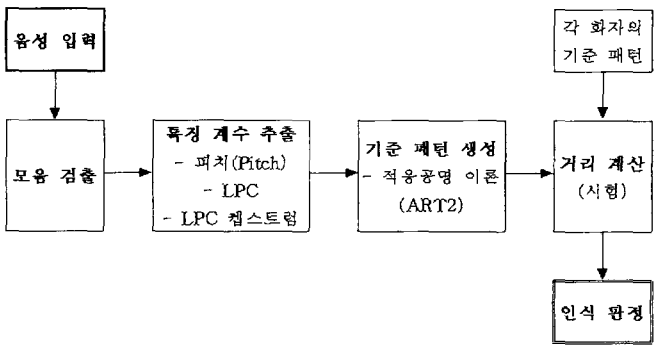
$$z_{ji} = \frac{u_i}{1-d}$$

⑮ 승리한 F_2 유니트로부터 나오는 상의하달 방식의 가중치들을 변경한다.

$$z_{ij} = \frac{u_i}{1-d}$$

그런 다음, 입력 벡터를 제거하고, 모든 비활성 F_2 유니트들을 복구한다. 새로운 입력 패턴과 함께 1단계부터 다시 시작한다.

본 연구에서는 ART2 모델에서 입력 패턴들을 성공적으로 다루기 위해서, F_1 계층을 순방향 연결과 역방향 연결을 모두 포함하는 6개의 부계층으로 나누었다. F_1 상의 모든 부계층들은 동일한 수의 유니트들(4개)을 갖는다. 따라서, 계층들은 F_2로의 상향 연결과 F_2로부터의 하향 연결을 제외하고는 완전하게 상호 연결되지 않는다.



[그림 2] 텍스트-독립 화자인식 시스템의 구성도

[Fig. 2] Block diagram of text-independent speaker recognition system

4. 실험결과 및 고찰

성도의 성문 임피던스는 -12dB/oct, 방사 임피던스는 6dB/oct의 기울기가 생기며, 결국 -6dB/oct의 경사를 갖는 저역통과 필터의 특성을 갖는다. 따라서 6dB/oct의 기울기를 갖는 고역통과 필터로 고역을 강조하여 평탄한 특성을 얻을 필요가 있다. 이렇게 함으로써 높은 주파수 영역에서 신호대 잡음비(SNR)가 저하되는 것을 보상할 수 있다. 이와 같이 고역 강조 필터를 통과한 신호에 대해서 음성신호 이외의 성분(잡음)을 제거하기 위해서 5KHz의 저역통과 필터를 사용한다.

4.1 피치 검출

계산량이 적은 시간 영역의 파라메터만을 사용해서 실시간으로 음성 구간을 검출하고 주변 환경의 변화에 따라 음성을 검출하는 파라메터의 임계값을 적응시키는 방법을 통해서 주변 환경의 영향을 적게 받도록 설계하였다. 한 분석 구간의 단시간 에너지는 다음과 같이 정의할 수 있다.

$$E = \sum_{i=0}^{N} |x(i)|$$

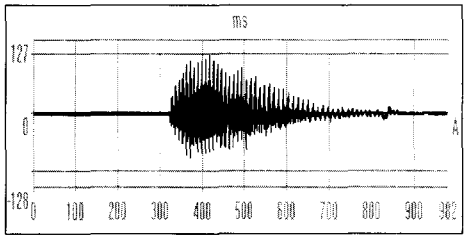
이 값이 에너지 상한값(TU)을 넘으면, 음성이 존재한다고 간주한다. 유성음 특히 모음의 경우 이 값이 크기 때문에 단시간 에너지만 갖고도 음성의 끝점을 검출할 수 있지만, 이 값이 에너지 상한값보다 작으면 에너지 하한값(TL)보다 큰가를 확인한 후, 에너지 하한값보다 크면 바로 이전 구간이나 바로 이후 구간이 유음성 구간인가 묵음 구간인가에 따라 결정된다. 그런데 무성음의 경우 거의 묵음 정도의 에너지 값을 가지므로, 무성음이 유성음보다 큰 값을 가지는 측정 방법을 추가로 사용하게 된다.

이와 같은 처리 과정을 통해 검출된 순수한 음성 부분만을 가지고 피치를 검출하게 되는데, 본 연구에서는 음성의 저주파수 성분만을 가지고 피치 주기를 검출하는 단순화 역필터 추정법(Simplified Inverse Filter Tracking Algorithm)을 사용하였는데, 음성 신호의 주기 성분이 대개 1.25KHz 이하인 저주파 대역에 존재하므로 표본화율을 2.5KHz 정도로 하게 되면 주기 정보를 함유하면서 단순하고 안정된 정보를 얻을 수 있기 때문이다.

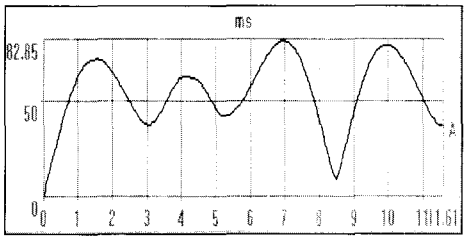
4.2 음성특징 추출

화자인식 실험에 사용할 특징 파라메터 추출 방법은 선형예측 계수(LPC)와 선형예측 계수들로부터 재구성된 선형 예측 켑스트럼 계수(LPCC), 그리고 모음 부분에서 검출한 피치 주기를 계수화하여 적용하고자 한다.

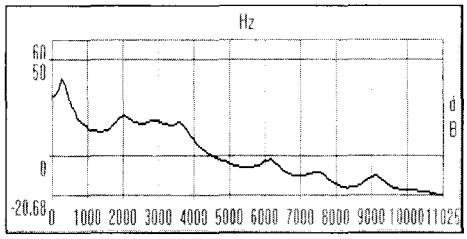
선형예측 계수와 선형예측 켑스트럼 계수의 경우 13차의 계수로 구하였는데, 대개 계수의 차수는 샘플링율과 밀접한 관련이 있으며, 여기에 채널 특성 등을 고려하여 2~5차를 추가한다. 본 연구에서는 화자의 특징을 가장 잘 나타내주는 특징 계수를 찾아내어 이를 신경망의 입력 패턴으로 사용하기 위해서, 3가지 종류의 특징 파라메터 값을 사용하였다.



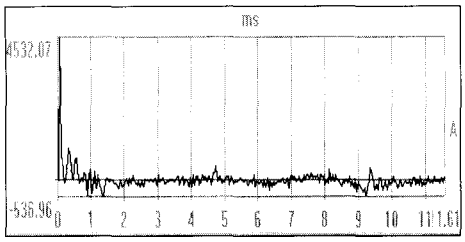
(a) '일'(/il/)의 파형(남성 화자)



(b) Pitch 주기



(c) LPC



(d) LPC Cepstrum

[그림 3] '일'(/il/)의 파형(남성 화자) 의 예
[Fig. 3] Example of waveform for /il/

4.3 신경망과 화자인식의 적용

피치 계수를 입력 벡터로 사용하는 경우 하나의 입력 벡터는 1X40 개의 원소로 표시되며, 선형예측 계수 및 선형예측 켑스트럼 계수를 입력 계수로 사용하는 경우에는 하나의 입력 벡터가 13X1개의 원소로 구성된다. 은닉층에서 노드의 수는 특별히 정해진 선택 기준 없이 대부분 신경망 설계자에 의해 임의로 정해지고 있다. 본 연구에서는 20개의 노드를 갖도록 하였으며, 출력층의 수는 적용하고자하는 신경망 원형에 따라 11개, 2개의 출력 노드를 갖도록 구성하여 실험하였다.

훈련을 위한 최대 실행 횟수를 1000회로 정하였고, 학습 시 목표 오차율은 0.01로 설정하였다. 신경망의 학습 속도는 통상 0.001~10의 범위 내에서 임의로 결정되는데, 이 값이 크면 학습이 빠르게 진행될 수 있지만 학습이 되지 않는 오차 최소점에 수렴하지 않을 수 있다. 반면에 너무 작으면 오차가 적어지는 형태로 학습이 이루어져서 최종적으로 오차 최소점에 도달되지만 각 학습 단계에서의 연결 강도 변화량이 너무 작아서 전체 학습 시간이 매우 길어지게 된다. 따라서, 본 연구에서는 0.1로 정하여 실험하였다.

4.4 실험 결과

화자들이 발음한 숫자음에서 추출한 피치를 입력 벡터로 사용하여 신경망을 훈련한 후, 임의의 숫자음 발음을 제공한 임의의 화자를 결정하는 실험을 하였다. <표 2>는 피치 계수를 적용한 텍스트-독립 화자 식별 실험을 적응 공명 이론 신경망 모델을 이용한 결과표이다. <표 2>에 나타난 것처럼 각 화자별 평균 인식률을 평균하여 95.14%의 전체 평균 인식률을 얻었다. 또한 숫자음들 중에서 '일'을 제공한 화자 A의 인식률이 48.4%로 가장 저조하였다.

<표 1> 신경망 파라메터 결정을 위한 예비 실험 결과
<Table 1> Simulation result for the determination of Neural network parameters

은닉층 노드수	학습율	목표 오차	실행 횟수			인 식 률		
			Pitch	LPC	LPCC	Pitch	LPC	LPCC
5	0.01	0.01	107	101	101	98.61	98.20	98.58
	0.1		58	56	55	98.96	98.20	98.60
	1		19	17	17	98.96	98.69	98.82
10	0.01	0.01	104	97	97	99.04	98.87	99.20
	0.1		55	52	51	99.06	98.87	99.20
	1		18	15	14	99.21	98.97	99.25
20	0.01	0.01	117	102	101	99.41	99.15	99.37
	0.1		61	56	55	99.50	99.19	99.40
	1		30	26	17	99.87	99.52	99.90

<표 2> 피치 계수와 적응공명 이론 신경망 모델을 적용한 텍스트-독립 화자인식 실험 결과

<Table 2> Simulation result of text-independent speaker recognition using pitch and ART2 neural network model

화자번호 / 입력패턴	A	B	C	D	E	F	G	H	I	J
PT00	90.16	97.43	98.98	65.10	98.06	97.67	97.37	98.28	79.76	97.57
PT01	48.40	96.09	98.95	99.52	99.21	97.69	96.05	98.78	95.11	98.08
PT02	91.74	72.83	99.33	98.48	93.25	98.18	97.79	98.86	90.19	95.20
PT03	94.94	95.80	98.92	98.59	96.85	97.65	97.56	98.91	97.98	98.48
PT04	82.60	77.46	99.07	99.35	98.88	87.94	99.00	98.74	98.34	98.67
PT05	96.25	98.31	99.54	98.86	97.97	92.48	92.94	99.76	99.61	95.19
PT06	93.54	95.95	99.23	96.97	99.06	97.63	97.77	98.88	97.57	98.59
PT07	55.19	95.72	99.41	98.89	97.05	97.34	96.53	98.82	97.83	98.08
PT08	95.02	94.34	99.25	99.13	99.08	97.47	98.63	98.60	98.09	98.62
PT09	99.46	76.54	98.56	99.17	98.12	93.52	98.06	98.74	94.99	97.79
평균인식률	84.73	90.05	99.12	95.41	97.75	95.79	97.17	98.84	94.95	97.63

적응공명 이론 신경망 모델을 사용하여 피치, 선형예측 계수, 선형예측 켑스트럼 계수의 세 가지 방법으로 텍스트-독립 화자인식 실험을 한 결과는 <표 3>과 같다.

<표 3> 적응공명 이론 신경망 모델을 사용한 텍스트-독립 화자인식 실험 결과

<Table 3> Simulation result of text-independent speaker recognition using ART2 neural network model

특징계수 / 인식률	피치	선형예측 계수	선형예측 켑스트럼 계수
최고 인식률	99.12	97.70	98.99
최저 인식률	84.73	80.60	87.34
평균 인식률	95.14	93.47	94.46

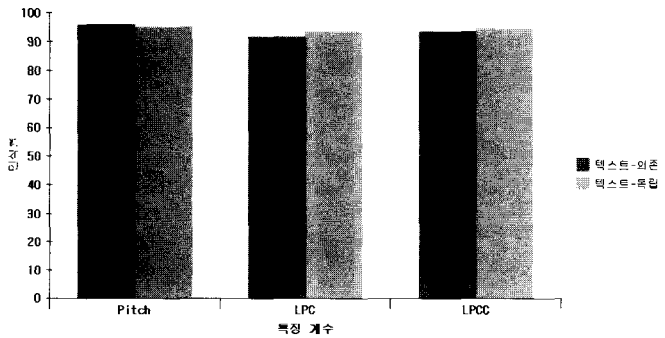
실험 결과에 의하면, 피치를 사용하여 신경망 모델로 실험한 결과가 가장 인식률이 높음을 알 수 있다. 따라서, 피치를 사용하면 다른 계수들을 사용하는 것보다 0.68~1.6% 이상의 인식률을 높일 수 있다. 또한 적응공명 이론 신경망 모델을 사용하여 피치, 선형예측 계수, 선형예측 켑스트럼 계수의 세 가지 방법으로 텍스트-의존 화자인식 실험을 한 결과와 비교한 결과를 [그림 4]에서 보여 주고 있다.

실험 결과를 분석해 보면, 텍스트-독립이나 텍스트-의존 모두 피치를 사용하는 것이 인식률이 높게 나타남을 알 수 있으며, 특히 텍스트-의존의 경우 가장 높은 인식률을 보이고 있다.

<표 4> 적응공명 이론 신경망 모델을 사용한 화자인식 실험 결과

<Table 4> Simulation result of speaker recognition using ART2 model

특징계수 / 인식률	피 치		선형예측 계수		선형예측 켑스트럼 계수	
	텍스트 독립	텍스트 의존	텍스트 독립	텍스트 의존	텍스트 독립	텍스트 의존
최고 인식률	99.12	99.84	97.70	98.91	98.99	99.46
최저 인식률	84.73	91.13	80.60	65.74	87.34	73.56
평균 인식률	95.14	95.82	93.47	91.57	94.46	93.37



[그림 4] 화자인식 실험 결과 비교

[Fig. 4] Simulation result of speaker recognition

5. 결론

지금까지 화자인식 연구는, 음성에 포함되어 있는 개인차가 선천적, 후천적인 영향 이외에 발성하는 텍스트 내용에도 의존하기 때문에 주로 텍스트 의존 방법으로 연구되어 왔으나, 텍스트 독립 화자인식을 위해서는 미리 등록한 단어가 아닌 경우에도 화자를 인식할 수 있도록 특징 파라메터를 개발하고, 실용화가 가능하도록 처리 방법을 간략화 해야할 것이다. 이를 위해서, 화자가 발성한 음성에서 모음을 검출하여 화자인식에 사용하는 방법을 제안하였다.

본 연구에서 화자인식을 위해 적응공명 이론 신경망 모델을 사용하였으며, 신경망의 입력패턴으로 사용될 특징 파라메터로 피치 주기, 선형예측 분석, 선형예측 켑스트럼 분석의 세 가지 특징 파라메터 추출 방법을 적용하여 비교하였다.

실험 결과 피치를 사용하는 것이 텍스트-독립의 경우 95.14%, 텍스트-의존의 경우 95.82%의 인식률을 보임으로, 다른 선형예측 계수들을 사용하는 것보다 인식률이 높음을 알 수 있었다. 따라서 화자인식의 경우 텍스트-독립, 의존 방법 모두 피치를 사용하는 것이 인식률을 높이는데 도움이 될 것으로 사료된다.

앞으로 연구과제는 최고 인식률과 최저 인식률 사이의 격차를 줄여서 평균 인식률을 높여야 할 것이며, 파라메터를 적용하는 과정을 개선한다면, 실시간 화자인식 시스템의 구현 및 각종 제품에 응용 될 수 있을 것으로 기대된다.

※ 참고문헌

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

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

[3] 中田和男, 音聲情報處理の基礎, オ-ム社, 1981.

[4] J. D. Markel & A. H. GRAY Jo., Linear Prediction of Speech, Springer Ver-lang, N.Y., 1976

[5] 鈴木久喜, 音聲の 線形豫測, コロナ社, 1978.

[6] 김 형래, 김 연숙, 배 성근, "음석 신호에 있어서 포만트 영향을 제거한 국부 봉우리와 골에 의한 피치 검출법", 대한전자공학회 하계 학술대회 논문지 제18권 제1호, pp. 297-301, Apr. 1981.

[7] B. Gold and L. R. Rabbiner, "Parallel Processing Techniques for Estimating Pitch Periods of speech in Time Domain", J. Acoust. Soc. Am., Vol.2, No.2, Pt.2, pp.442-448, Aug, 1965.

[8] Sadaoki Furui, "Digital Speech Processing, Synthesis, and Recognition", Marcel Dekker, Inc., pp.81-85, 1992.

[9] Y. H. Pao, "Adaptive Pattern Recognition and Neural Networks", Addison Weseley, 1988.

[10] S. Y. Kung, Digital Neural Networks, Prentice-Hall, pp.78-85, 1993.

[11] J. A. Freeman and D. M. Skapura, "Neural networks; Algorithms, Applications, and Programming Techniques", Addison Wesley Pub., 1991.

김에녹

1977년 광운대학교 전자계산학과 졸업

1982년 건국대학교 대학원 전자공학과(공학석사)

1993년 건국대학교 대학원 전자공학과(공학박사)

1983년~현재 연암공업대학 컴퓨터정보기술과 교수

관심분야 : 디지털 신호처리, 화자 인식, 화상인식 컴퓨터 네트워크, Fuzzy System