論文2001-38SP-5-12

새로운 시간축 정규화 방법을 이용한 한국어 고립단어 인식기

(Korean isolated word recognizer using new time alignment method of speech signal)

南 明 祐 *, 朴 奎 洪 *, 魯 承 容 *

(Myung Woo Nam and Kyu-Hong Park, and Seung Ryong Rho)

요 약

본 논문에서는 음성신호의 발성길이와 상관없이 일정한 크기의 파라미터를 얻을 수 있는 새로운 방법을 제안하였다. 음성인식기의 성능은 음성신호에서 추출된 파라미터간의 유사도(패턴간의 거리)를 어떻게 비교하는지에 따라 결정된다. 그러나 화자에 따른 음성신호의 변이나 발성속도의 차이는 음성신호에서 일정한 크기의 파라미터 추출을 어렵게 한다. 제안한 방법은 음성신호에서 얻어진 파라미터를 스펙토그램의 형태로 표현한 뒤 2차원 DCT(Discrete Cosine Transform)를 이용해 일정한 크기의 파라미터로 정규화시키는 방법이다. 제안한 방법의 유효성을 입증하기 위해 청각세포를 모델링한 32개의 대역통과 필터로부터 얻어진 음성신호의 파라미터를 2차원 DCT 방법으로 가공한 후, 신경 회로망의 입력으로 사용하였다. 또한 기존 방법과의 인식률 비교를 위해 기존의 정규화된 입력을 구하는 방법 중 하나를 선택하여 비교 실험을 수행하였다. 실험결과 제안한 방법은 기존 방법에 비해 화자종속 및 화자독립 고립단어 인식에서 더 높은 인식률과 빠른 인식속도를 얻을 수 있었다.

Abstract

This paper suggests new method to get fixed size parameter from different length of voice signals. The efficiency of speech recognizer is determined by how to compare the similarity(distance of each pattern) of the parameter from voice signal. But the variation of voice signal and the difference of speech speed make it difficult to extract the fixed size parameter from the voice signal. The method suggested in this paper is to normalize the parameter at fixed size by using the 2 dimension DCT(Discrete Cosine Transform) after representing the parameter by spectrogram. To prove validity of the suggested method, parameter extracted from 32 auditory filter-bank(it estimates auditory nerve firing probabilities) is used for the input of neural network after being processed by 2 dimension DCT. And to compare with conventional methods, we used one of convetional mehhods which solve time alignment problem. The result shows more efficient performance and faster recognition speed in the speaker dependent and independent isolated word recognition than conventional method.

I. 서 론

최근 연구수준에서 머물러 있던 음성관련 기술들이 속속 상품화되면서 다양한 제품들이 선보이고 있다. 실용화되고 있는 음성신호처리 분야로는 사람의 음성을 인식시켜 명령을 처리하게 하는 음성인식을 비롯해 통신분야에서 많이 사용되는 음성코딩, 사람의 음성과 유사한 소리를 인공적으로 만드는 음성합성, 사람음성에 포함된 개별적 특성을 보안에 이용하는 화자인식 등이 있다. 그러나 음성신호처리는 음성에 포함된 많은 불규칙적인 동적 특성들로 인해 아직까지도 매우 어려운

* 正會員, 서울市立大學校 電子工學科
(CAD&VLSI Lab., Dept. of Elec. Eng., The University Of Seoul)

接受日字:2000年12月27日, 수정완료일:2001年5月16日

과제로 알려져 있다.[4]

현재 음성인식기에 주로 사용되는 방법으로는 HMM (Hidden Markov Model)과 신경 회로망(Neural Networks), 그리고 DTW(Dynamic Time Warping) 등이 있는데, 이러한 방법들은 연속음성인식과 고립단어 인식에서 좋은 결과들을 보여주고 있다. 특히 연속음성인식의 경우는 그 응용분야가 광범위하여 활발한 연구가 진행되고 있으며, 저가의 응용제품도 많이 개발되고 있다. 그러나 연속음성인식기들은 높은 인식률을 얻기 위하여 복잡한 알고리즘으로 구현되기 때문에 하드웨어 구현시 회로의 복잡도가 증가하는 단점이 있다. 따라서 이러한 연속음성인식 방법들은 자동통역기와 같은 대규모 시스템에 적용함이 유리하다고 생각된다. 그러나 음성인식이 이러한 고급 시스템에만 이용되는 것이 아니라 이동전화나 PDA(Personal Digital Assistant), 가전제품 등에도 사용될 수 있으므로 높은 인식률을 가지면서 빠른 인식속도와 간단한 구조를 갖는 고립단어 인식기의 필요성도 매우 높다. 특히 고립단어 인식기는 그 응용분야의 특성 때문에 숫자음 또는 간단한 명령어 인식 위주로 연구가 진행되고 있다.[5]

신경 회로망을 이용한 음성인식은 소규모 고립단어 인식에서 높은 성능을 보여주나 음성신호의 불규칙성과 시간정보의 표현, 그리고 처리시간 등의 문제점들로 인해 실제 응용에는 많은 제약을 받고 있다.[4] 특히 화자에 따른 음성신호의 변이나 발성 속도 차이를 정규화시키기 위해 신경 회로망 자체에 시간적 개념을 부여한다는 것이 어렵기 때문에 신경 회로망을 이용한 음성인식은 대부분 발성길이가 짧은 고립단어 인식이나 분할된 음소인식에 주로 이용되고 있는 실정이다.

기존의 고립단어인식기들은 음성신호의 발성길이를 정규화시키기 위하여 DTW, TDNN(Time Delay Neural Network), LBG 알고리즘 등을 사용하였다. 이러한 방법들은 음성신호로부터 얻어진 파라미터들을 반복적인 작업을 통해 몇 개의 대표적인 패턴들로 수렴시키거나 가장 유사한 경로를 찾아가는 알고리즘들이다.[5,7,9] 따라서 많은 연산시간을 필요로 하며 음성신호 전체의 특징들을 모두 보존하는데 한계가 있다.

본 논문에서는 고립단어로부터 얻어진 파라미터를 고립단어의 발성길이와 상관없이 일정한 크기로 정규화시킬 수 있는 새로운 방법을 제안하였다. 그리고 제안한 방법을 신경 회로망과 결합하여 유효성을 입증하였다. 먼저 청각세포를 모델링한 32개의 대역통과 필터에 음성신호를 통과시킨 후, 여기서 얻어진 스펙트럼을 2차원 DCT에 적용, 고정 크기의 파라미터로 변환하였다. 그리고 얻어진 고정 크기의 파라미터를 다층구조 신경 회로망에 입력으로 사용하였다.

2차원 DCT를 이용하여 얻은 고정 크기의 특징벡터는 신경 회로망의 가장 큰 단점인 신호의 시간적 변화 정보를 효율적으로 정규화시켜 준다. 또한 청각모델을 통해서 얻어진 스펙트럼의 정보를 큰 손실없이 높은 비율로 압축하여 준다. 따라서 기존의 방법들에 비하여 상대적으로 적은 양의 데이터를 사용하게 되므로 빠른 인식속도를 얻을 수 있었으며, 인식실험에서도 높은 인식률을 얻을 수 있었다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서는 음성신호의 끝점검출 방법과 청각모델을 이용한 파라미터 추출방법에 대해 설명하였다. Ⅲ장에서는 2차원 DCT의 정의와 이것을 이용하여 정규화된 파라미터를 얻는 방법에 대해 설명하였다. 그리고 Ⅳ장에서는 인식에 사용된 신경 회로망의 구조에 대해 기술하였고, Ⅴ장은 제안한 방법을 이용하여 실시한 다양한 실험 결과를 나타내었다. 마지막으로 결론은 Ⅵ장에서 기술하였다.

Ⅱ. 특징 파라미터 추출

1. 음성신호 검출

일반적으로 잡음은 음성신호의 변화에 비해 시간적으로 매우 느리게 변화한다. 따라서 잡음에 대한 신호를 사전에 미리 선형 예측함으로써 잡음이 섞인 음성신호를 분리해낼 수 있다. 본 논문에서는 이러한 원리를 이용하여 잡음으로부터 음성신호를 검출해내는 방법을 사용하였다. 사용된 음성신호 검출방법은 잡음신호에 대해 6차 선형예측계수를 구한 후 이러한 계수를 가지는 필터에 음성신호를 통과시켜 얻은 잔차신호(선형예측 에러신호)를 이용한다. 이때 얻어진 잔차신호는 잡음의 잔차신호와 잡음이 제거된 음성신호의 합과 같다. 음성신호의 검출을 위해 사용된 에너지 함수는 다음과 같다.

$$E = \sqrt{E_{s_1} * E_{s_2}} - \sqrt{E_{n_1} * E_{n_2}} \qquad (1)$$

여기서 E_{s_1}은 입력 음성신호의 에너지이며 E_{s_2}는 음성신호에 대한 잔차신호의 에너지, E_{n_1}은 잡음신호의

에너지, 그리고 E_{n_2}는 잡음신호에 대한 잔차신호의 에너지이다.

2. 청각모델을 이용한 특징추출

견고한 음성신호의 특징추출 방법은 추후 단계의 인식률을 좌우하는 중요한 연구분야이다. 과거에는 음성신호의 특징추출 방법으로 선형예측계수(Liner Prediction Coefficient)나 켑스트럼계수(Cepstrum Coefficient)를 많이 사용하였는데, 연구결과 켑스트럼계수들이 음성인식 부분에 있어 다른 선형예측계수들에 비해 더 좋은 성능을 보인다고 밝혀졌다.[4]

최근에는 청각세포의 주파수 응답특성을 이용한 다양한 음성신호의 특징추출 방법들이 제시되고 있다. 청각세포는 낮은 주파수 신호의 변화에 대해서는 높은 해상도로 반응하는 반면, 높은 주파수 신호의 변화에 대해서는 낮은 해상도로 반응하는 특징이 있다. 청각세포의 주파수 응답특성을 이용한 청각모델은 다양한 음성신호의 상태(잡음, 스트레스, 마이크로폰 변이와 채널 왜곡) 변화에 대해서는 편차가 적지만, 서로 다른 소리에 대해서는 특징 파라미터의 편차가 크게 나타나는 것으로 알려져 있다.[1,2]

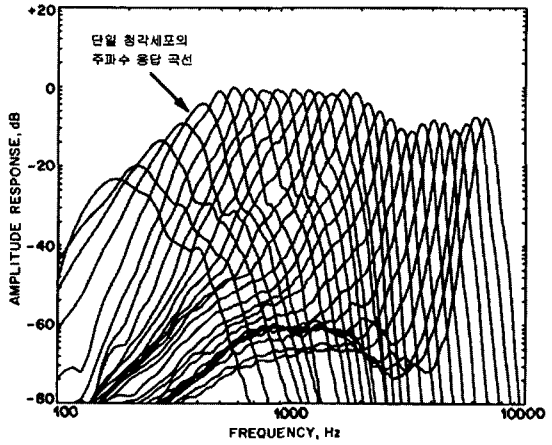


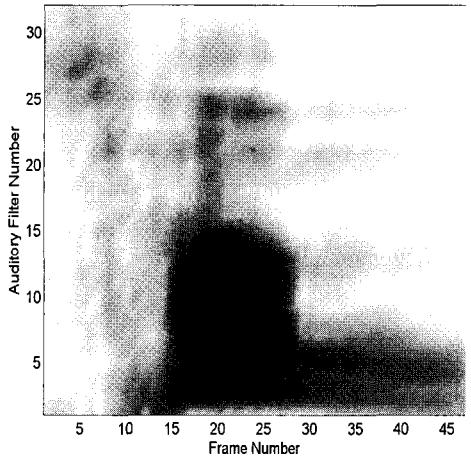
그림 1. 청각세포의 주파수 응답

Fig. 1. Frequency response curves of basilar membrane.

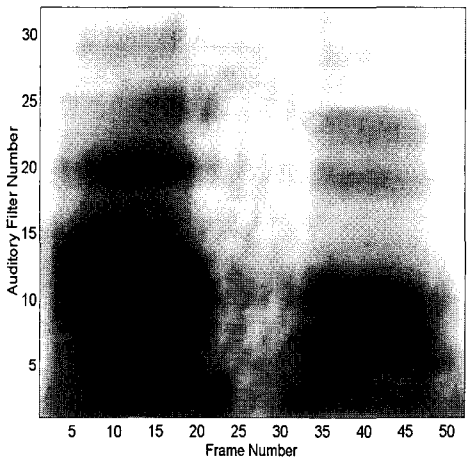
본 논문에서는 인간의 청각세포를 32개의 대역통과 필터로 모델링하였으며, 사용된 필터의 주파수 특성은 그림 1과 유사하도록 5차 IIR(Infinite Impulse Response) 필터로 설계하였다. 저주파에서 고주파 쪽으로 갈수록 필터의 대역폭이 넓어지게 설계하였으며, 저주파 쪽에 보다 많은 필터들이 할당되게 구현하였다. 사용된 청각모델은 하드웨어로의 구현시 비교적 간단하며, 일반적인 FFT(Fast Fourier Transform)에 비하여 적은 계산량과 잡음에 강한 특성을 가진다.

Ⅲ. 2차원 DCT를 이용한 음성정보의 압축

청각모델을 사용하여 구한 정보를 시간적 흐름(frame number)에 따라 배열하면 2차원적인 모양이 되며 스펙트로그램(spectrogram)의 형태를 보이게 된다. 그림 2에서는 숫자음 '삼'과 '아홉'으로부터 32개의 청각모델 필터를 사용하여 얻은 정보를 2차원적으로 보여주고 있다.



(a) 삼 (SAM)

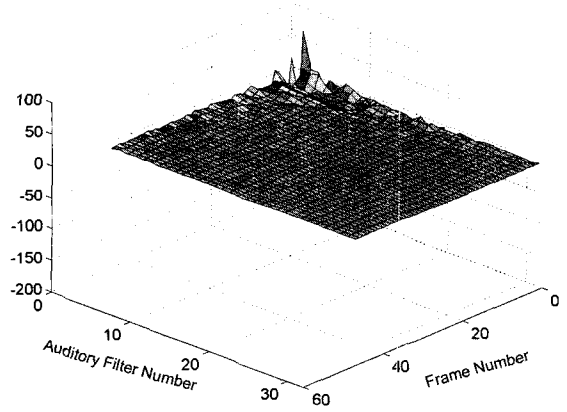


(b) 아홉 (A-Hop)

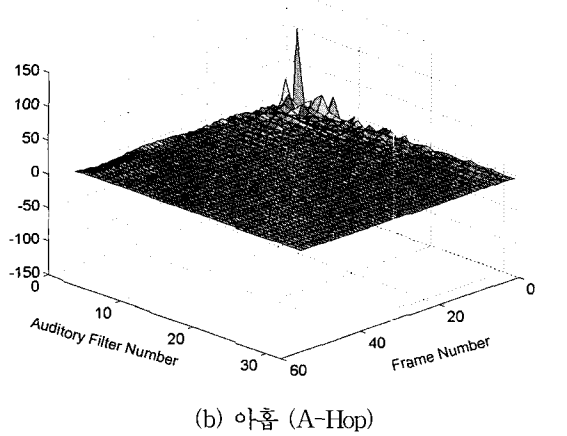
그림 2. 청각모델을 통하여 얻은 음성신호의 특징 스펙트럼

Fig. 2. Feature spectrogram by auditory model.

고립단어에 대한 청각모델의 2차원 정보는 2차원 DCT를 사용하여 그림 3과 같이 변환된다. 이때 얻어진 DCT 계수는 필터축과 프레임축의 원점 주위에서는 큰 값을 가지는 반면, 원점에서 멀어질수록 점점 작은 값을 가지게 된다. 따라서 원점에서 멀리 떨어진 계수는 제거되어도 원래의 2차원 정보에는 큰 영향을 미치지 않게 된다. 이러한 특성은 가변 프레임 수를 가지는 음성신호를 일정한 크기로 정규화시키는데 사용될 수 있다. 즉, 얻어진 DCT계수에서 전체정보에 영향을 미치지 않는 부분들을 제거하면서 다른 단어들과 같은 크기로 파라미터들을 정규화시키는 것이다. 그림 4는 얻어진 DCT 계수를 일정한 크기(11×6)로 잘라낸 후, 다시 처음의 2차원 정보로 복원시킨 경우이다. 그림 4에서 알 수 있듯이 처음 정보에서 미세한 부분들만이 제거되었을 뿐 전체적인 정보는 그대로 보존되고 있다.

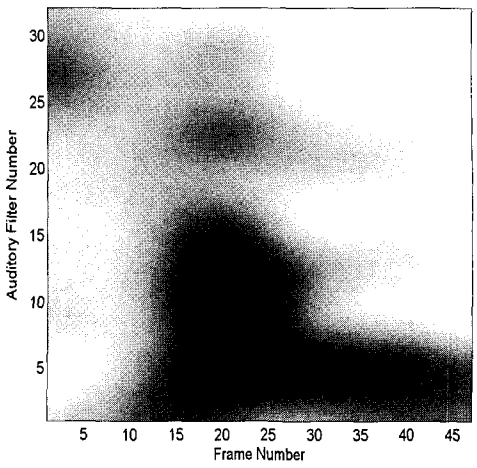


(a) 삼 (SAM)

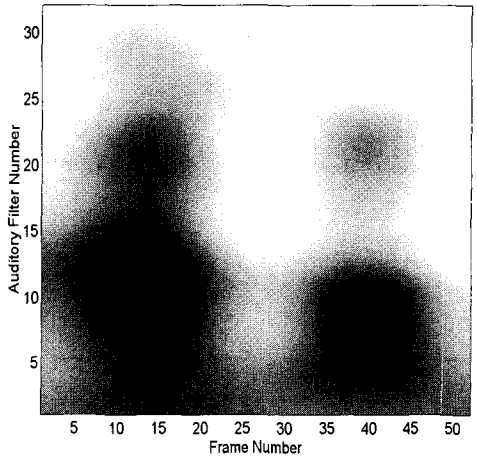


(b) 아홉 (A-Hop)

그림 3. 음성신호의 특징정보를 압축한 2차원 DCT 계수
Fig. 3. Extracted 2D-DCT coefficients by auditory model.



(a) 삼 (SAM)



(b) 아홉 (A-Hop)

그림 4. 압축된 2차원 DCT계수로부터 복원한 특징 스펙트럼
Fig. 4. Restructed feature spectrogram from 2D-DCT coefficients.

2차원 DCT 계수를 구하는 방법은 다음과 같다.

$$f(u,v)=\frac{2}{N}C(u)C(v)\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}d(m,n)\cos\frac{(2m+1)u\Pi}{2N}\cos\frac{(2n+1)v\Pi}{2N} \quad (2)$$

식 (2)에서 $d(m,n)$는 m, n번째 청각모델을 사용하여 추출된 특징의 크기값이며, $f(u,v)$는 u, v번째 2차원 DCT값이다. $C(u), C(v)$는 식 (3)과 같다.

$$C(u),C(v)=\begin{cases}\sqrt{\frac{1}{2}} & \text{for } u=0, v=0\\ 1 & \textit{otherwise}\end{cases} \quad (3)$$

2차원 DCT는 각각의 행에 대해 1차원 DCT를 한 다음, 각각의 열에 대해 1차원 DCT를 한 것과 같다. 따라서 순차적으로 각 프레임별 1차원 DCT 계수를 구한 다음, 음성신호의 종료와 동시에 프레임 방향으로 1차원 DCT를 수행하여 2차원 DCT 계수를 얻게 된다. 구해진 2차원 DCT 계수 중에서 평균에너지와 관계되는 $f(0,0)$ 계수에 대해서는 음성신호의 크기에 따라 크게 변하므로 특징추출 파라미터에서 제외하였다.

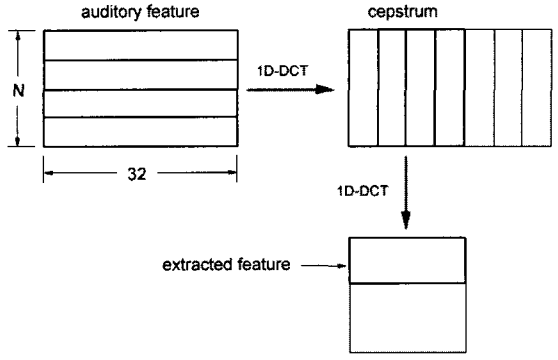


그림 5. 2차원 DCT를 이용한 특징 추출과정
Fig. 5. Feature extraction from speech signal using 2D-DCT.

본 논문에서는 2차원 DCT의 특성을 이용하여 음성신호로부터 얻어진 다양한 길이의 파라미터들을 일정한 크기의 파라미터들로 정규화시키는데 사용하였다. 제안한 방법은 몇 가지의 장점을 가지고 있다. 먼저 기존의 방법들에 비해 빠르게 정규화를 수행할 수 있다. 2차원 DCT는 이미 영상처리분야에서 많이 사용되고 있는 일반화된 방법이다. 따라서 영상처리분야에서 개발된 고속 알고리즘이나 하드웨어를 사용하게 된다면 실제 응용에서 큰 이점을 얻을 수 있을 것이다. 다음으로 정규화과정 중 파라미터들의 원래정보가 거의 그대로 보존된다는 것이다. 따라서 파라미터들의 유사도를 비교할 때 각 단어들이 가진 좀더 다양한 정보들이 사용될 수 있다. 또한 DCT는 파라미터들을 높은 비율로 압축하여 준다. 때문에 작은 DCT 계수만 가지고도 원래의 정보를 보존할 수 있다. 이는 고립단어들을 작은 크기의 파라미터로 큰 정보의 손실없이 정규화할 수 있다는 의미가 된다. 그러므로 파라미터들의 유사도 비교에 사용되는 전체 연산량이 크게 줄어들게 된다.

제안한 방법의 유효성을 입증하기 위하여 음성에서 얻어진 파라미터를 2차원 DCT를 이용해 11×6의 고정된 크기의 DCT 계수로 변환한 후, 신경 회로망의 입력으로 사용하였다.

Ⅳ. 다층구조 신경 회로망

다층구조 신경 회로망은 대표적인 정적(static) 신경 회로망으로서 교사학습에 의한 인식이나 분류를 수행한다. 현재 음성인식에 사용되는 많은 신경 회로망 모델들은 다층구조 신경 회로망을 기반으로 하여 음성에 포함된 많은 변이를 수용하고자 시도하고 있다. 기본적인 다층구조 신경 회로망은 음성인식의 가장 기본적 문제인 시간 정렬(time alignment) 능력이 부족한 단점이 있다.[4] 대부분의 다층구조 신경 회로망에서는 고정된 입력 뉴우런 수만큼의 패턴을 입력하여 각 출력 뉴우런의 단어(class)에 해당하는 출력값을 구하는 정적 패턴 인식(static pattern recognition) 기능만을 수행하게 된다. 따라서 음성을 인식하기 위해서는 길이가 다른 음성을 정규화한 후 전체 음성을 모두 입력으로 사용하여 출력값을 얻든지, 몇 개의 프레임만을 입력으로 하여 출력값들을 차례로 구한 후 이들을 누적하여 최적의 단어를 고르는 방법을 사용한다. 본 논문에서는 전자의 방법을 사용하여 인식실험을 수행하였다. 다층구조 신경 회로망을 이용한 음성인식은 간단한 구조로 인하여 하드웨어 구현이 용이하고 소규모 고립단어 인식에서 높은 성능을 보여주는 장점 때문에 많이 이용되고 있다. 그러나 인식하고자 하는 어휘수에 비례하여 신경 회로망의 크기가 증가하므로 대 어휘 단어인식 등에 적용이 어렵다는 단점이 있다. 따라서 다층구조 신경 회로망은 대부분 발성길이가 짧은 구분단어 인식이나 분할된 음소인식에 주로 이용되고 있은 실정이다.

신경 회로망의 학습방법에는 한 노드가 인접 노드의 활성화에 기여한다면 이들 두 노드를 연결하는 연결선의 연결 강도는 증가되어야 한다는 hebbian rule에 의한 방법과 실제 출력된 값과 제시된 값간의 차이를 줄이도록 노드간의 연결 강도를 조절하고, 상위 오차를 아래층으로 전파시켜 하위층이 오류를 교정하는 delta rule에 의한 방법이 있다. 그리고 학습자료의 제시 여부에 따라 희망출력패턴이 입력과 함께 제시되는 교사학습(supervised learning)과 그렇지 않은 자율학습(unsupervised learning)으로 나눌 수 있다.

실험에 사용된 다층구조 신경 회로망은 delta rule 방법으로 학습되었으며 교사학습 방법을 사용하였다. 2차원 DCT를 이용하여 얻은 파라미터는 일정한 크기로 절단되어 정규화된 후 신경 회로망의 입력으로 사용됐다. 사용된 신경 회로망은 66개의 입력 뉴우런과 100개의 뉴우런으로 구성된 1개의 중간층, 그리고 51개의 출력노드를 갖도록 설계하였다. 전체적인 고립단어 인식기의 구성은 다음과 같다.

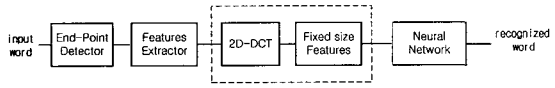


그림 6. 제안한 고립단어 인식기의 구성도
Fig. 6. Block diagram of proposed isolated word recognizer.

V. 실험 결과

실험의 음성 데이터는 ETRI의 음성 데이터베이스를 사용하였다. 남성 20명과 여성 20명이 4회 발성한 20개의 숫자음('일', '이', '삼', '사', '오', '육', '칠', '팔', '구', '영', '하나', '둘', '셋', '넷', '다섯', '여섯', '일곱', '여덟', '아홉', '공')과 남성 13명이 2회 발음한 31개의 고립단어인 총 4,006개의 음성 데이터를 이용하여 인식 실험을 수행하였다. 사용된 단어들은 16kHz의 표본화율과 16bit로 저장된 음성 데이터들이다.

신경 회로망은 먼저 화자종속 고립단어 인식실험을 위해 숫자음에서는 남성 20명과 여성 20명이 2회 발성한 20개의 단어를, 고립단어에서는 남성 13명이 1회 발성한 31개의 단어를 사용하여 훈련을 수행하였다. 다음으로 화자독립 고립단어 인식실험을 위해 숫자음에서는 남성 10명과 여성 10명이 4회 발성한 20개의 단어를 사용하였으며, 고립단어에서는 남성 6명이 2회 발성한 31개의 단어를 사용하여 훈련을 수행하였다. 그리고 나머지 음성 데이터들을 이용하여 고립단어와 숫자음 인식실험을 수행하였다. 모든 음성신호는 0.015초 간격으로 절단한 후 50%씩 중복시켜 32개 청각모델 필터를 이용해 파라미터를 추출하였으며, Hamming 창함수를 사용하였다. 추출된 2차원 DCT 계수 중에서 평균에너지와 관계되는 $f(0,0)$ 계수에 대해서는 음성신호의 크기에 따라 크게 변하므로 특징추출 파라미터에서 제외하였다(모두 0으로 대체하였다).

제안한 방법의 유효성을 입증하기 위해 기존의 정규화된 입력을 구하는 방법 중 하나를 선택하여 비교 실험을 수행하였다. 선택된 방법은 다음과 같다. 먼저 시간 순서대로 나열된 인접하는 프레임간 Euclidean 거리를 누적시켜 구한 후 일정하게 분할한다. 그리고 각각의 분할된 구간들에서 중간 프레임을 대표 패턴(Vector Quantization)으로 추출하는 방법이다.[9]

표 1에서는 제안한 방법과 기존 방법을 사용한 화자종속 인식률을 비교하였다. 그리고 표 2에서는 제안한 방법과 기존 방법을 사용한 화자독립 인식률을 비교하였다. 표 1과 2에서 알 수 있듯이 숫자음의 경우는 제안한 방법과 기존 방법이 거의 유사한 인식률을 보여주나 고립단어의 경우 긴 음절을 가지는 단어들이 존재하기 때문에 기존 방법으로는 높은 인식률을 얻을 수 없었다.

표 1. 화자종속 고립단어에 대한 인식률
Table 1. Recognition rate of speaker dependent isolated words.

학습단어 / 학습방법	화자종속 고립단어 인식률 (%)			
	숫자음		고립단어	**전체**
	남성	여성		
NN with 2D-DCT (11×6)	98.37	98	96.52	**97.63**
NN with VQ (32×9)	97.5	97	88.83	**94.44**

표 2. 화자독립 고립단어에 대한 인식률
Table 2. Recognition rate of speaker independent isolated words.

학습단어 / 학습방법	화자독립 고립단어 인식률 (%)			
	숫자음		고립단어	**전체**
	남성	여성		
NN with 2D-DCT (11×6)	97.87	95.62	92.85	**95.45**
NN with VQ (32×9)	96.87	96.25	79.95	**91.02**

기존 방법을 사용한 신경 회로망은 입력 프레임의 개수가 증가함에 따라 인식률도 비례하여 증가하는 결과를 보여준다. 그러나 고립단어들은 단음절에서부터 여러 음절까지 다양한 단어들이 존재하므로 신경 회로망 입력을 위한 프레임의 개수를 계속 늘릴 수가 없는 단점이 있다. 이런 이유로 숫자음과 고립단어에서 최고

표 3. $f(0,0)$ 계수(평균 에너지)에 따른 인식률
Table 3. Recognition rate by $f(0,0)$ coefficient.

학습단어 / 학습방법	화자독립 고립단어 인식률 (%)						전체	
	숫자음				고립단어			
	남성		여성					
	$f(0,0)=0$	$f(0,0)\neq0$	$f(0,0)=0$	$f(0,0)\neq0$	$f(0,0)=0$	$f(0,0)\neq0$	$f(0,0)=0$	$f(0,0)\neq0$
NN with 2D-DCT (11×6)	97.87	94.15	95.62	94	92.85	94.47	**95.45**	**94.19**

인식률을 얻을 수 있었던 프레임의 수는 많은 차이를 보였으며 전체 음성 데이터를 하나의 고정된 프레임 수를 사용해서 실험했을 경우 최고 91.02% 정도의 낮은 인식률을 얻었다. 그러나 본 논문에서 제안한 방법은 단어의 음절수에 따른 인식률의 변화가 크지 않아 안정된 인식률을 얻을 수 있었다.

그림 7과 8에서는 DCT 계수에서 추출되는 파라미터의 크기 변화에 따른 신경 회로망의 인식률을 나타내었다. 그리고 그림 9에서는 기존 방법(VQ)을 이용했을 경우 입력 프레임변화에 따른 신경 회로망의 인식률 변화를 보였다. 2차원 DCT로 파라미터를 정규화할 경우 화자종속 및 화자독립 음성인식에서 11×6개의 특징 파라미터를 사용하는 방법이 저장공간의 용량 및 인식률에서 상대적으로 좋은 결과를 보여주고 있다.

제안한 방법은 추출되는 파라미터의 크기와 인식률이 일정구간까지 비례하다가 다시 반비례하는 결과를 보여주고 있다. 이러한 이유는 2차원 DCT로부터 고정된 크기의 파라미터를 취하는 과정에서 각 단어의 개별특성들이 제거되기 때문이다. 그러나 DCT에서 추출되는 파라미터의 크기가 커짐에 따라 다시 개별특성들이 나타나게 되고 오인식의 원인으로 작용되어 인식률의 저하가 발생한다. 따라서 높은 인식률을 얻기 위해서는 적절한 크기의 파라미터를 추출하는 것이 중요하다.

본 논문에서 제안한 방법은 2차원 DCT를 이용하여 작은 크기의 파라미터로부터 좋은 인식률을 얻을 수 있다. 따라서 기존 방법들과 동일한 크기의 파라미터를 사용할 경우 제안한 방법은 매우 좋은 결과를 얻을 수 있으며, 50개 이상의 고립단어에 대해서도 좋은 결과를 보여주었다.

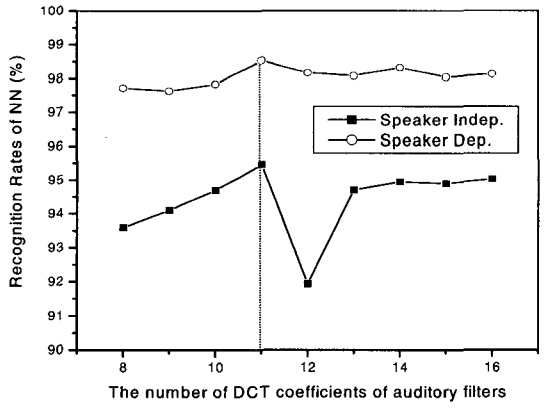


그림 7. 필터축의 DCT 계수 변화에 따른 신경 회로망의 인식률
Fig. 7. Recognition rate from change of DCT coefficients of Auditory filter.

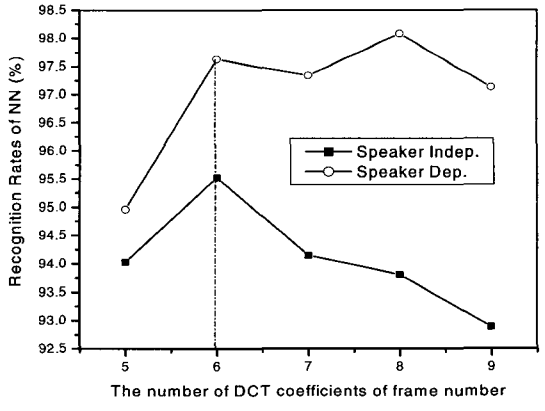


그림 8. 프레임축의 DCT 계수 변화에 따른 인식률
Fig. 8. Recognition rate from change of DCT coefficients of frame number.

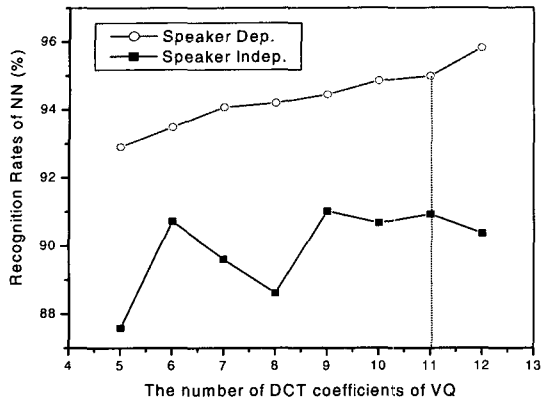


그림 9. 기존 방법(VQ)의 프레임 변화에 따른 신경 회로망의 인식률

Fig. 9. Recognition rate from change of frame numbers of VQ.

VI. 결 론

본 논문은 새로운 시간축 정규화 방법을 이용하여 얻어진 파라미터를 신경 회로망의 입력으로 사용하여 한국어 숫자음과 고립단어 인식 실험을 수행하였다. 먼저 인간의 청각세포를 모델링한 32개의 대역통과 필터를 사용하여 1차 파라미터를 구한 후, 2차원 DCT를 이용하여 2차 파라미터를 구하였다. 그리고 2차 파라미터를 일정한 크기로 잘라내어 미리 학습된 신경 회로망의 입력으로 사용하였다. 실험 결과 제안한 방법은 화자종속과 화자독립 모두에서 기존 방식보다 높은 인식률을 얻을 수 있었다. 2차 파라미터의 경우 11×6의 크기로 정규화하는 것이 저장공간 및 인식률에서 가장 좋은 결과를 보여주었다. 그리고 기존의 고립단어 인식기에 비해 시스템의 복잡도가 크게 줄었으며, 빠른 인식속도와 높은 인식률을 얻을 수 있었다.

제안한 방법을 하드웨어로 구현할 경우, 적은 메모리와 간단한 회로만으로도 고립단어 인식기의 구현이 가능하며, 고립단어 인식을 필요로 하는 가전제품이나 PDA 또는 이동전화기 등의 다양한 부분에 응용이 가능할 것이다.

참 고 문 헌

[1] Ghitza, O., "Auditory models and human performance in tasks related to speech coding and speech recognition", Speech and Audio Processing, IEEE Transactions on, vol 2, issue 1, part 2, pp. 115-132, Jan. 1994.

[2] J.L. Goldstein, "Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering", Hearing Res., vol. 49, pp.33-60, 1990.

[3] G. K. Wallace, "The JPEG still picture compression standard", IEEE Trans. Consumer Electron. vol 38 no.1, pp.18-34, Feb. 1992.

[4] 오영환, 음성언어정보처리, 홍릉과학출판사, 1997.

[5] Jianping Huang, Anthony Kuh, "A neural network isolated word recognition system for moderate sized databases", Neural Networks, IEEE International Conference on, vol. 1, pp. 387-391, 1993.

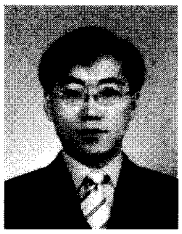
[6] M.T. Hagan, H.B. Demuth, M. Beale, Neural network design (PWS Publishing Company. 1996).

[7] Figueiredo, F.L.; Violaro, F., "An isolated word speech recognition system based on Kohonen neural network", Neural Networks, 1998. Proceedings. Vth Brazilian Symposium on, pp. 151-156, 1998.

[8] Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of speech recognition (Prentice-Hall International, Inc. 1993).

[9] Shan Zhu; Dao Wen Chen; Tai Yi Huang, "Feature parameter curve method for high performance NN-based speech recognition", Acoustics, Speech, and Signal Processing, ICASSP-96. Conference Proceedings., IEEE International Conference on, vol 1, pp. 1-4, 1996.

저 자 소 개

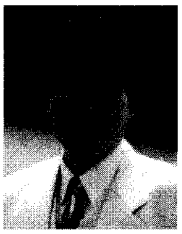


南 明 祐(正會員)

1969년 4월 25일생. 1992년 2월 : 서울시립대학교 제어계측공학과 졸업(학사). 1994년 2월 : 동 대학원 전자공학과 졸업(석사). 1999년 2월 : 동 대학원 박사과정수료. 현재 동 대학원 박사과정. 주관심분야 : 음성인식, 음성신호처리

魯 承 容(正會員)

1944년 5월 1일생. 1971년 2월 : 한양대학교 전자공학과 졸업. 1988년 2월 : 한양대학교 대학원 전자공학과(공학박사). 1982년~현재 : 서울시립대학교 전자·전기공학부 교수. 주관심분야 : 음성신호처리, VLSI 회로설계



朴 奎 洪(正會員)

1974년 4월 8일생. 1997년 : 서울시립대학교 전자공학과 졸업. 1999년 : 동 대학원 전자공학과 졸업(석사). 2001년 : 동 대학원 박사과정수료. 현재 동 대학원 박사과정. 주관심분야 : 음성인식, 음성신호처리