DHMM과 신경망에서 숫자음 인식률 비교

박정환* · 이원일* · 황태문** · 이종혁*

*경성대학교 컴퓨터공학과

**대진정보전자공업고등학교

Digit Recognition Rate Comparision in DHMM and Neural Network

Jung-hwan Park* · Won-il Lee* · Tae-moon Hoang** · Jong-hyeok Lee*

* Kyungsung University

** Deajin Information Electric Technology Highschool

E-mail : one-point@hanmail.net

요 약

음성 신호는 언어정보, 개인성, 감정 등의 여러 가지 정보를 포함한 음향학적인 신호인 동시에 가장 자연스럽고 널리 쓰이는 의사소통 수단의 하나이다. 본 연구에서는 저장된 음성 신호에서 추출한 특징 파라미터를 사용한 경우와 음성 특징파라미터에 입술 패턴에 대한 영상정보를 동시에 사용한 경우 DHMM과 신경망을 통하여 각각 인식률을 비교해 보았다. 그 결과 입술패턴에 대한 영상정보도 음성인식에 사용 할 수 있음을 알 수 있었다.

키워드

MFCC, 신경망, HMM, DHMM

I. 서 론

음성 신호는 언어 정보, 개인성, 감정 등의 여러 가지 정보를 포함한 음향학적인 신호인 동시에 음성은 가장 자연스럽고 널리 쓰이는 의사소통 수단의 하나이다.

언어와 더불어 음성은 인간이 지니고 있는 정보 전달 수단 중 누구나 쉽게 사용할 수 있다는 뛰어난 장점을 지니고 있으며 전달 속도면에 있어서나 말하는 사람과 듣는 사람 모두 다른 작업을 병행하면서 자유롭게 정보를 주고 받을 수 있다는 점에서도 커다란 이점이 있다.

음성은 개개인의 발음속도와 발음지속시간 및 성도의 차이에 따라 음성 패턴에 대하여 비선형적인 변동이 발생한다. 그러므로 이러한 변동의 제거, 시간축의 정규화 및 조음결합 문제는 음성인식 연구에 중요한 과제로 제시되고 있다.[1]

본 연구에서 구현 및 실험을 위한 데이터로는 숫자음(공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구)을 사용하며 음성 특징 파라미터는 12차 MFCC계수로 하고 영상 특징 파라미터로 16단계 히스토그램을 사용하여 신경망과 DHMM을 통하여 각각의 인식률을 비교한다.

II. 음성과 영상

2.1 음성정보분석

음성신호를 표현하기 위해서는 에너지, 영교차율과 같은 기본적인 특징으로부터 LPC, 필터뱅크 모델과 같은 복잡한 표현 방식에 이르기까지 다양한 특징들이 제안되어 왔다.

필터뱅크 모델은 구현이 용이하며 잡음이나 다른 형태의 왜곡 현상에 강하기 때문에 현재 많이 사용된다.

본 실험에서 사용된 음성 특징 파라미터는 MFCC를 통해 구한 12차의 특징 값과 여기에 영교차율과 에너지 값을 추가한 14개의 값을 사용하였고, 영상 특징 파라미터는 입술영상의 16차 히스토그램을 사용하였다.

가청주파수 범위는 20Khz 정도이므로 음향의 경우 샘플링 주파수는 그의 2배인 40Khz 정도이면 된다. 그러나 음성의 경우 3.4Khz 정도의 대역폭만 가지면 알아듣는데 지장이 없으므로 음성특징파라미터를 추출하기 위하여 음성은 샘플링 주파수 8Khz에 16bits 양자화 하여 저장하였다.[2]

이렇게 저장된 데이터들 중 실제 음성구간만을 추출하기 위하여 전처리 과정을 거쳤다. 그림 1은 음성특징추출을 위한 전처리 과정이다.

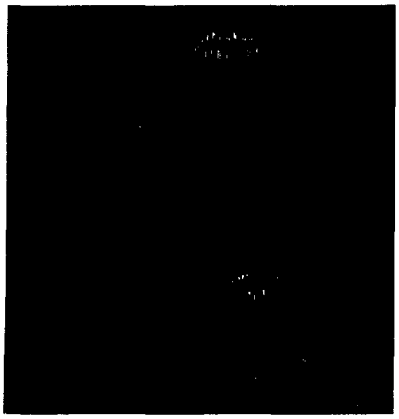

그림 1. 전처리과정

코드북(codebook)이란 음성신호로부터 추출한 훈련벡터(training vector)들간의 거리값의 평균왜곡이 최소가 되도록 설정한 대표벡터의 집합이다.

코드북을 작성하기 위한 방법으로 본 연구에서 LBG 알고리즘을 사용하였다.

2.2 영상정보분석

영상처리에서 가장 간단하면서 유용한 도구 중의 하나가 히스토그램이다. 히스토그램은 영상의 명도 내용을 요약한 것이라 할 수 있다.

한 이미지에서 밝은 점과 어두운 점이 분포할 때 그 분포의 범위와 값을 표현한 것이다. 이를 그래프로 나타낸 것을 히스토그램 그래프라고 한다. 히스토그램은 영상에 대한 상당한 정보를 가지고 있고 계산하기에 간편하므로 여러 영상처리에 이용되고 있다.[3] 그림 2는 입술 영상의 gray이미지와 히스토그램의 예를 보여준다.

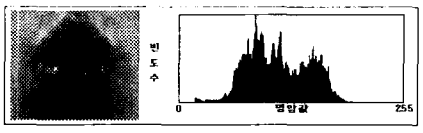

그림 2. gray이미지와 히스토그램

III. 신경망

1943년 McColloch와 Pitts가 처음 도입한 신경망이란 상호 연결된 많은 수의 인공 뉴런들을 이용하여 생물학적 시스템의 계산 능력을 모방하는 소프트웨어나 하드웨어로 구현된 계산 모형을 말한다.[4]

인간의 뇌는 뉴런이라는 기본 신경 단위로 구성되어 있으며, 이들 상호간의 연결 형태에 따라 지식을 암호화하거나 해독하게 된다. 신경망에서는 생물학적인 뉴런의 기능을 단순화시킨 인공뉴런(또는 노드)을 사용하며 이들은 연결을 통해 상호 연결되어 인간의 인지 작용이나 학습과정을 수행하게 된다.

다층 신경망은 입력층과 출력층 사이에 하나이상의 계층을 갖는 신경망을 말한다. 은닉층이란 명칭은 출력층과 같이 목표 출력 값이 학습 시에 알려지지 않고 숨겨져 있다는 의미에서 붙여진 이름이다. 다층 신경망의 구조 중 대표적인, 한 개의 은닉층을 갖는 3층 신경망의 예를 그림 3에 나타내었다.[2][5]

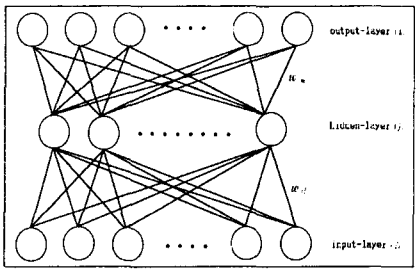

그림 3. 신경망의 구조

IV. HMM

HMM은 관찰 확률과 관찰되지 않는 상태 천이 확률을 이용하여 시간에 따라 변화하는 음성 신호의 특징을 잘 표현할 수 있는 이중적인 통계과정으로 현재 널리 사용되는 중요한 음성인식 기법으로 소규모의 인식 시스템에서부터 대규모의 인식 시스템까지 모두 적용이 가능하며, 화자종속과 화자독립에 관계없이 적용될 수 있고, 인식 성능 또한 우수하다.

HMM을 이용한 음성의 모델링은 음성으로부터 추출한 특징 파라미터를 관측열로 직접 모델링하는 연속 출력 확률 분포 HMM(CHMM)과 추출한 특징 파라미터를 벡터양자화 과정을 거쳐 코드북의 코드워드 인덱스로 매핑한 관측열을 출력확률로써 모델링하는 이산 출력 분포 HMM(DHMM)으로 구분할 수 있다.

N개의 상태를 갖는 HMM은 아래와 같이 표시된다.

$$\lambda = (A, B, \Pi) \quad (1)$$

λ : HMM 모델
A : 상태 천이 확률
B : 관찰 확률
Π : 초기 상태 확률

상태 천이 확률 A는 각 상태 간의 천이 확률을 나타내는 행렬이다. 시간 t에서 상태 s_t가 i일때 시간 $t+1$에서의 상태 s_{t+1}이 j로 천이할 확률 a_{ij}로서 나타내며 N개의 상태를 갖는 HMM의 경우 상태 천이 확률 A는 N×N의 행렬이 된다.

$A=a_{ij}$: 상태 s_t로부터 상태 s_{t+1} 로의 천이 확률

$$a_{ij}=P(s_{t+1}=j|s_t=i) \quad (2)$$

$B=b_j(o_t)$: 상태 s_t에서 상태 s_{t+1}로 천이할 때 o_t 을 출력하는 확률

$$b_j(o_t)=P(o_t|s_t=j) \quad (3)$$

초기 상태 확률 Π 는 1부터 N까지 각각의 상태에서 t=1인 첫 번째 관찰이 일어날 확률을 나타내는 확률이다.

$\Pi=\pi_i$: π는 초기상태 확률의 집합

$$\pi_i=P(s_i=i) \quad (4)$$

DHMM은 입력신호의 D차원으로 이루어진 특징을 L개의 코드워드로 구성된 코드북에 의해 양자화한다. 이때 코드북의 각 코드워드들은 특징 공간의 한 곳을 불연속적으로 나타낸다.[6]

VI. 구현과 실험

실험에 사용된 음성 및 영상 데이터는 숫자음 데이터로 "공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구" 10가지이고, 화자 10명(20대 남자7명, 여자3명)이 숫자음 발성시 동시에 영상과 음성을 저장하였다. 음성 특징 파라미터를 추출하기 위하여 음성은 샘플링 주파수 8Khz에 16bit 양자화하여 저장하였다.

본 연구에서는 인식을 위한 인식기로 신경망과 DHMM을 이용하였다. 신경망은 음성에서 구분되어진 초성, 중성, 종성 데이터를 이용하여 그림 4와 같이 구성하였다. 초성, 중성, 종성의 신경망은 음성에서 12차의 MFCC계수, 에너지와 영교차율, 그리고 영상에서 입술모양에 대한 16차의 히스토그램을 사용하였다.

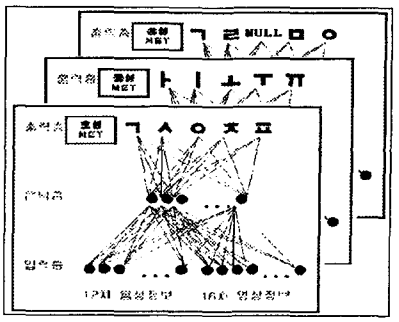

그림 4. 초성, 중성 및 종성 인식 신경망의 구성

DHMM에서는 음성 계수를 VQ하여 코드북을 구성하고, 이산적인 유한개의 코드북 벡터 번호를 심벌로 하여 관측확률을 계산하기 위해, 상태 i에서 j로의 천이확률 a_{ij} 및 그 천이에서 심벌 k의 관측확률 $b_{ij}(k)$를 Baum-Welch 알고리즘을 이용하여 계산한다. 그림 5는 하나의 DHMM을 훈련시키는 과정을 도식화한 것이다.

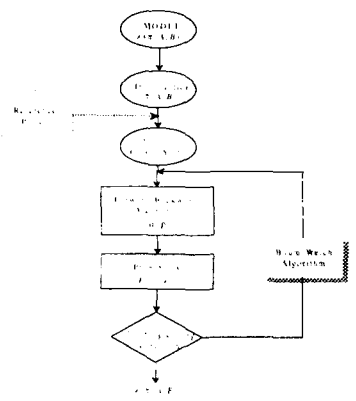

그림 5. DHMM Training Diagram

DHMM에 음성 또는 음성과 영상을 입력하였을 때 state 수에 따른 인식결과를 그림 6에 나타내었다. state 4에서 전체(학습과 평가)에서 인식률이 좋았으며 음성만을 사용한 인식보다 음성과 영상을 통한 인식결과가 향상됨을 알 수 있었다.

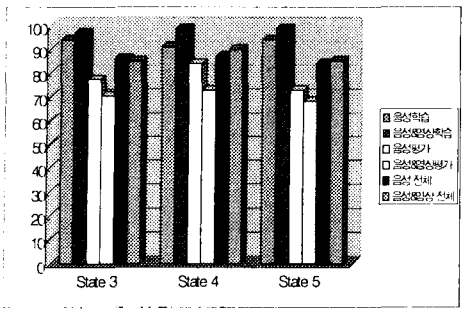

그림 6. DHMM의 State 변화에 따른 인식 결과 비교

DHMM을 이용한 학습화자와 평가화자의 비율에 따른 인식 결과를 그림 7에 나타내었다. 학습 80%, 평가 20%에서 좋은 결과를 얻을 수 있었다. 여기서 데이터를 충분히 확보하여야 인식률이 향상됨을 알 수 있다.

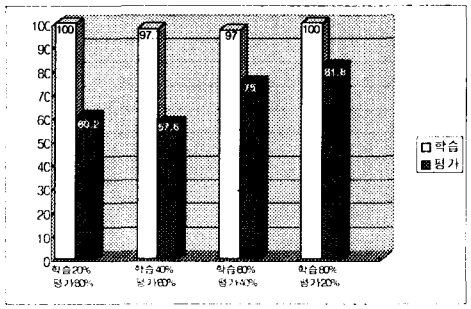

그림 7. DHMM을 이용한 학습화자와 평가화자의 비율에 따른 인식 결과

음성과 영상의 가중치에 따른 실험결과를 그림 8에 나타내었다. 음성가중치를 영상가중치보다 높게 했을 때 인식결과가 향상됨을 알 수 있었다. 이것은 영상이 인식의 보조수단임을 알 수 있다.

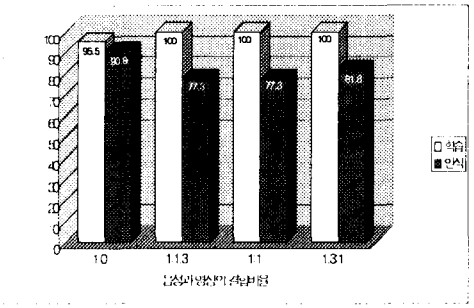

그림 8. 음성과 영상의 가중치에 따른 인식 결과

신경망을 이용한 숫자음 인식 결과를 그림 9에 나타내었다. 신경망에서도 음성과 영상정보를 동시에 사용하였을 경우 좋은 인식률을 얻을 수 있었다.

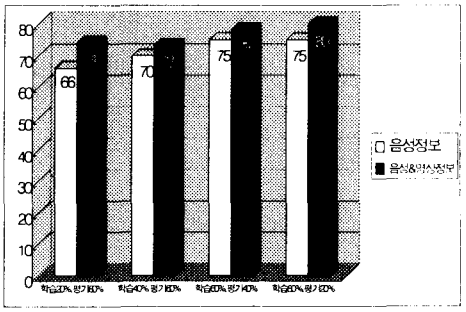

그림 9. 신경망을 이용한 학습화자와 평가화자의 비율에 따른 인식 결과

V. 결 론

본 연구에서는 음성 정보만을 이용한 기존의 시스템에 가시적인 영상 정보를 추가하여 보다 안정되고, 높은 인식률을 보이는 숫자음 인식 시스템을 제안하였다.

실험 결과 DHMM과 신경망 모두 가시적인 정보가 보다 높은 인식률을 가지며 음성 인식 시스템의 설계에 있어서 상당히 유용한 특징 파라미터로 사용될 수 있음을 확인할 수 있었다.

그러나 저장된 영상의 입술 패턴의 기울어짐, 거리와 조명등에 따라 서로 다른 형태의 출력값이 나오므로 영상 정보의 특징을 추출하는데 어려움이 많았다. 이러한 문제는 앞으로 지속적인 연구를 통해 해결해 나가야 할 것이다.

참고문헌

[1] 백인찬, 권영헌, 이건상, 김형관, 남호성, 양성일, "필터뱅크 출력의 개수 변화에 따른 TDNN과 MSVQ 음성인식기에 관한 연구", 한국음향학회지, 제16권, 제 7호, pp31~32, 1997

[2] 조현욱, "음성과 영상 정보를 이용한 우리말 숫자음 인식", pp15~20, 2002.

[3] 백준기, "첨단 영상 미디어 서비스와 영상 복원기술", 전자공학회지, Vol.23 No.6, pp.28-39, 1996.

[4] 이재령, "TDNN을 이용한 한국어 숫자음 인식", pp20~22, 1999

[5] 이상원, "음성정보에 얼굴 영상정보를 추가한 음성인식에 관한 연구", pp9~12,1998

[6] 양태영, "한국어 연결 숫자음 인식의 성능 향상을 위한 알고리즘",pp10~11, 2000