# 음성과 영상정보를 이용한 우리말 숫자음 인식

조현욱* · 이종혁*

*경성대학교 컴퓨터공학전공

# Digit Recognition using Speech and Image Information

Hyun-wook Cho* · Jong-hyeok Lee*

* Kyungsung University Computer Eng.

E-mail : hw@bus114.com

## 요 약

본 논문에서는 음성에서 얻어지는 특징 파라메타와 음성을 발성할 시 얻을 수 있는 가시적 데이터에서 추출되는 파라메타를 함께 이용하여 우리말 숫자음 인식을 시도하였다. 실험에서는 음성정보만을 이용한 기존의 방법과 영상정보의 추가할 경우의 인식성능을 비교, 검토하였다. 전체에서 50%를 학습시켰을 경우 학습된 화자의 경우 100%, 학습되지 않은 경우에는 78%의 인식률을 보였다.

## ABSTRACT

We propose The Korean digit recognition system using speech and image information. In the experiments, we investigate that image information affect recognition rate. Recognition rate of learned data and testing data show 100%, 78% each other.

## 키워드

음성인식, 신경회로망, MFCC

## Ⅰ. 서 론

음성신호는 언어정보, 개인성, 감정 등의 여러 가지 정보를 포함한 음향학적인 신호이다. 뿐만 아니라 음성은 가장 자연스럽고 널리 쓰이는 의사소통 수단의 하나이다. 이로 인하여 기계와의 의사소통을 위한 다양한 방법들 중에서 인간의 발성 음으로부터 자동으로 언어 정보를 추출하는 음성인식에 관한 연구가 근래에 활발히 수행되고 있다.

기존의 음성인식 방법은 대부분 음성 그 자체에 대한 특징 파라메타를 구하여 이것을 음성인식 모델들의 입력으로 사용하여 인식을 시도하였다. 이러한 방식은 입력 레벨, 배경잡음, 다른 화자에 따른 차이 등의 외부요인으로 인해 인식결과에 많은 영향을 받아 왔다. 하지만 인간은 음성을 인식하기 위하여 청각뿐만 아니라 이를 지원하기 위하여 자신의 시각을 사용한다. 가시적인 신호는 주의력을 집중하도록 보조해줄 뿐만 아니라 청취자가 음향적인 음성을 이해하기 곤란할 때 유용한 정보자원을 제공하기도 한다. 특히, 잡음이 많은 환경을 접할 경우, 인간이 음성을 인식하는데 있어서 가시적인 정보는 중요한 역할을 한다. [1]

따라서 이렇게 저하되는 인식률을 높여 보고자, 본 논문에서는 종래에 이용되던 음성에서 얻어지는 특징 파라메타와 음성을 발성할 시 얻을 수 있는 가시적 데이터에서 추출되는 파라메타를 함께 이용하여 인식을 시도하였다. 실험 및 평가를 위한 데이터로는 단독 숫자 음을 사용하였으며 신경회로망인식기를 이용한 인식을 시도하였다. 이러한 실험을 통하여 음성정보만을 사용한 음성인식방법과 본 논문에서 실험한 영상 정보를 추가한 음성인식방법 두 가지 방법에 대한 인식성능을 비교 검토하였다.

## Ⅱ. 음성정보 분석

음성의 대표적인 특징 중의 하나는 다양성이다. 동일한 단어를 여러 사람이 발음하였을 경우 단어의 의미가 동일하더라도 음성 파형은 동일하지 않으며, 동일한 단어를 동일한 시간에 연속으로 발음하였다고 하여도 음성 파형은 동일하지

않다. 이와 같은 현상의 이유는 음성 파형에서는 음성의 의미정보 이외에도 화자의 음색, 감정상태 등과 같은 정보도 포함하고 있기 때문이다.

### 2.1 전 처리 과정

전 처리 과정(Pre-processing)은 음성신호를 본격적 분석과정에 들어가기 전에 처리하는 과정으로 환경적응, 끝점검출, 반향제거, 잡음제거 등이 있다. 잡음제거는 화자가 발음할 때 주변 환경이나 기계의 잡음을 효과적으로 제거하는 방법이며, 끝점검출은 입력된 디지털 음성신호로부터 음성인식에 필요한 음성구간만을 검출하는 처리과정으로서 일반적으로 에너지값과 영교차율( ZCR : Zero Crossing Rate)값을 이용하여 음성 구간을 검출한다.[2] 환경적응 기술은 주변 잡음환경이나 기기의 환경에 지능적으로 적응하는 기술이며, 반향제거 기술이란 음성의 반사로 인해 원래의 신호를 왜곡하는 현상을 제거하는 방법이다.

### 2.2 음성 특징 추출

사용자가 발음한 음성 신호는 디지털 처리를 통해 입력받게 되는데, 이 신호를 그대로 사용하기에는 정보량이 많고 불필요한 신호 요소들이 많이 포함되어 있기 때문에 인식 과정에 필요한 특징만을 추출하게 된다. 일반적으로 특징 추출에 이용되는 방법으로는 선형 예측 분석에 의한 LPC(Linear Prediction Coefficient) 추출법과 켑스트럼 추출 방식에 의한 MFCC(Mel Frequency Cepstrum Coefficients) 추출법 등이 있다.

음성은 비교적 짧은 구간에 대하여 조사해보면, 급격하게 변하지 않고 특정구간은 그 특성이 동일한 신호들이므로, 이전까지의 신호를 알면 이들 신호에 어떤 특정한 계수를 곱하여 이전 신호들을 더하면 현재의 신호를 알 수 있게 된다. 시간 t에서의 음성 샘플을 $x_t$라고 하고 현재의 음성 샘플을 이전 p 개의 샘플로부터 예측하고 이때, 예측값과 실제값간의 차이를 $e_t$라고 하면 다음 식이 성립한다.[3] 특정구간에서 $e_t$를 최소로 할 수 있는 $a_i$가 LPC계수가 된다.

$$x_t = -\sum_{i=1}^{p} a_i x_{t-i} + e_t \qquad (1)$$

켑스트럼 추출 방법 중 인간의 청각 인지과정을 고려한 방법이 MFCC추출 법이다. 인간의 청각은 저주파 부분에서 매우 예민하고 고주파 부분에서 덜 예민하다. 이러한 특성을 고려하여 주파수축을 재 변환한 것을 멜척도라 하고, 주파수(f)와의 관계는 다음과 같이 나타낼수 있다.

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (2)$$

MFCC는 FFT변환하여 구한 스펙트럼을 다시 멜 스케일로 나뉘어진 필터뱅크를 사용하여 $p$개의 필터뱅크출력을 구한 후 필터뱅크출력의 로그값에 DCT(Discrete Cosine Transform)을 적용함으로써 구할 수 있다. 다음은 M차의 MFCC를 얻기 위한 식이다.

$$c_i = \sum_{j=1}^{p} X_j \cos(\frac{i(j-0.5)\times\pi}{p}) \qquad (3)$$
$$(1 \le i \le M)$$

여기서 $C_i$는 $i$번째 MFCC를 p는 필터뱅크의 채널 수를 $X_j$는 $j$번째 필터뱅크출력의 로그값을 나타낸다.

## III. 영상정보분석

### 3.1 히스토그램

디지털 영상처리에서 가장 간단하면서 유용한 도구 중의 하나가 히스토그램(histogram)이다. 히스토그램은 영상의 명도도 내용을 요약한 것이라 할 수 있다.

빈약한 명암 값 분포를 가진 영상은 히스토그램 평활화라고 불리는 처리에 의해 향상될 수 있다. 히스토그램 평활화의 궁극적인 목적은 일정한 분포를 가진 히스토그램을 생성하는 것이다. 따라서 평활화를 수행한 히스토그램은 균일한 분포를 가질 수 있다. 그 결과 어두운 영상은 밝아지고 너무 밝은 영상은 조금 어두워져 적당한 명도 값을 유지하게 된다.

### 3.2 윤곽선 추출

윤곽선은 영상 안에서의 영역의 경계를 나타내는 특징으로 픽셀의 밝기의 불연속 점을 나타낸다. 윤곽선은 영상 안에 있는 물체의 윤곽에 대응되며 많은 정보를 가지고 있고 물체의 위치, 모양, 크기, 표면의 무늬 등에 대한 정보를 알려준다.[4]

윤곽선 검출 방법을 나눌 때 흔히 1차 미분법, 2차 미분법, 나머지로 나누고 있다. 1차 미분을 이용한 윤곽선 검출 방법에는 차분(Difference) 필터, Sobel 필터, Roberts 필터, Kirsch 필터, Robinson 필터, Prewitt 필터 등이 있고, 2차 미분을 이용한 에지 검출 방법의 대표적인 것은 라플라시안 윤곽선 필터(Laplacian edge filter)있다. 또한 이러한 연산자들을 이용한 응용이라고 볼 수 있는 캐니(Canny)연산자가 있는데, 이것은 먼저 가우시안 마스크를 이용하여 잡음을 제거한 후 윤곽선 검출 연산자를 수행하는 것이다.

음성을 발음하였을 시 입술모양과 케니연산자를 이용한 윤곽선 검출의 예는 그림1과 같다.

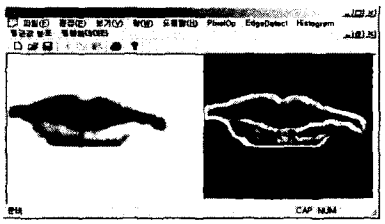

그림 1 케니연산자를 이용한 윤곽선 검출의 예

## Ⅳ. 신경회로망

신경회로망은 인간의 뇌의 구조를 컴퓨터로 구연한 것이다. 신경회로망은 인간의 두뇌처럼 예를 통한 학습, 일반화, 연상기억, 결함허용성 등의 특징이 있다. 이중 신경회로망의 가장 큰 장점인 학습은 패턴의 부류에 따라 가중치를 적당한 값으로 지정하는 과정이다. 즉, 원하는 패턴이 입력으로 주어질 때 지정된 노드가 일정한 출력값을 내도록 하는 것이다.

일반적으로 신경회로망이 커지고 복잡해질수록 더 나은 기능을 수행할 수 있게 된다. 신경회로망의 입력층과 출력층 사이에 새로운 층들을 추가하는 다층 신경회로망은 대체로 더 복잡한 기능을 수행할 수 있다. 다층 신경회로망에서 입력층(input layer)과 출력층(output layer) 사이의 층들은 은닉층(hidden layer) 혹은 중간층(internal layer) 이라 부른다.[5] 그림2는 하나의 은닉층을 가진 다층신경회로망의 예를 나타낸 것이다.

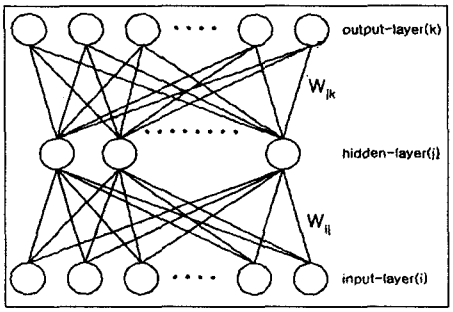


그림 2 다층 신경회로망의 예

## Ⅴ. 실험방법 및 결과

본 논문에서 제안한 음성인식 시스템의 블럭도는 그림 3과 같다.

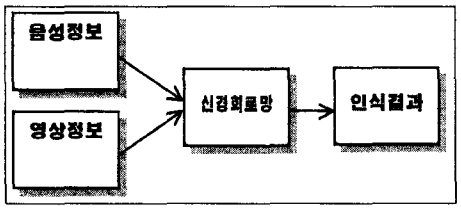


그림 3 음성인식 시스템의 블럭도

실험에 사용된 음성 및 영상데이타는 숫자음 데이터로 "공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구" 10가지 이고, 화자 10 명( 20대 남자7명, 여자3명)이 숫자음 발성시 동시에 영상과 음성을 저장하여 이중 5명(남자4명, 여자1명)의 데이터는 학습용으로, 나머지 5명(남자3명, 여자2명)의 데이터는 평가용으로 사용하였다.

### 5.1 특징 파라메타 추출

음성데이터를 얻기 위하여 샘플링 주파수 8KHz에 16bits/sample 양자화로 저장하였다. 저장된 데이터들을 실제 음성구간만을 뽑아내기 위하여 에너지 값과 영 교차율 값을 이용하여 시작점과 끝점을 검출하고 묵음구간을 제거하였다.

특징 파라메타를 추출하기 위하여 LPC와 MFCC 두 가지 방법을 사용하였다. 두 가지의 추출 방법들은 각각 12차의 특징 값이 나오도록 구성하고 실험을 통해 두 가지 방법의 특징 추출 성능을 비교하였다.

영상정보는 음성 저장과 동시에 화자의 발성시의 입술모형을 이용하여 256*256 크기의 Gray-Image로 초당 15프레임 저장하고 히스토그램 평활화를 통하여 명암 값이 균일한 이미지로 변환하였다.

이렇게 변환된 이미지를 캐니연산자를 사용하여 잡음에 강한 윤곽선을 추출하고 히스토그램을 이용하여 256개의 수치값으로 변환 하였다.

256개의 데이터는 실제 신경회로망의 입력데이터로 사용하기에는 상당히 많은 양이므로 각 8개의 연속된 데이터들의 평균값을 만들어 16개의 데이터로 줄여 인식을 위한 입력데이터로 사용하였다. 그림4는 입술모형과 이로부터 추출한 16차의 히스토그램을 나타낸 것이다.

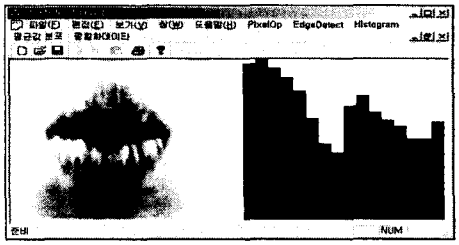


그림 4 입술모형과 16차 히스토그램

### 5.2 신경회로망의 구성

본 논문에서 사용한 숫자음 인식을 위한 신경회로망은 음성에서 세그멘트 되어진 초성, 중성, 종성 데이터를 위한 총 3개의 독립된 신경회로망을 구성하였다. 초성, 중성, 종성의 각각의 신경회로망은 입력 데이터로 음성에서 12차의 LPC 또는 MFCC계수, 에너지와 영 교차율, 그리고 영상에서 입술모양에 대한 16차의 히스토그램을 사용하여 총 30개의 값을 학습과 평가를 위한 입력값으로 사용하였다. 은닉층에는 10개의 노드를 두어 신경회로망을 보다 견고하게 하였고, 출력층은 각 회로망마다 서로 다른 목표 출력 값을 갖는데, 우선 초성NET는 "ㄱ, ㅅ, ㅇ, ㅊ, ㅍ" 총 5개의 초성들을 출력값으로 두고 중성NET는 "ㅏ, ㅣ, ㅗ, ㅜ, ㅠ" 총 5개의 값, 종성NET는 "ㄱ, ㄹ, ㅁ, ㅇ, null" 로 구성하여 5개의 출력값을 가지도록

신경회로망을 구성하였다.

### 5.3 실험 결과 및 검토

표 1은 인식시스템을 통해 기존에 학습되어진 화자의 데이터에 대한 인식률과 학습시키지 않은 평가용 화자의 데이터의 인식률을 비교한 것이다.

| 데이터의 종류 | 음성정보 | | 음성&영상정보 | |
|---|---|---|---|---|
| | LPC | MFCC | LPC | MFCC |
| 학습 | 47/50 | 50/50 | 50/50 | 50/50 |
| 평가 | 29/50 | 35/50 | 35/50 | 39/50 |

표 1 화자별 인식결과

학습된 화자에 대한 인식결과는 LPC를 음성파라메타로 사용한 경우에는 94%를 보였고, 나머지 모든 경우에서 100%의 인식률을 보였다. 평가용 화자의 결과를 살펴보면 음성정보만을 사용한 방법에서는 LPC를 이용하여 음성파라메타를 추출한 경우에는 58%의 인식률을 나타내고 MFCC를 이용한 경우에는 70%의 인식률을 나타내었다. 영상을 첨가한 방법에서는 LPC를 이용한 경우에 %의 인식률 MFCC를 이용한 경우에는 78%의 인식률을 보여 영상을 추가한 인식방법이 8% 정도의 인식률 증가가 있음을 알 수 있다. 또한 MFCC를 이용한 음성파라메타 추출방법이 LPC를 이용한 경우보다 평균10% 높은 인식률을 보이고 있다.

표 2는 평가용 데이터에 대한 초성, 중성, 종성의 각 음소별 인식결과를 나타낸 것이다.

| 데이터의 종류 | 음성정보 | | 음성&영상정보 | |
|---|---|---|---|---|
| | LPC | MFCC | LPC | MFCC |
| 초성 | 31/50 | 36/50 | 34/50 | 35/50 |
| 중성 | 32/50 | 38/50 | 40/50 | 44/50 |
| 종성 | 43/50 | 47/50 | 43/50 | 46/50 |

표 2 음소별 인식결과

음소별 인식결과에서 먼저, 종성의 인식률을 살펴보면 MFFC음성추출 방법의 경우 93% 정도의 좋은 결과를 보이는 것을 알 수 있고, 중성의 경우는 음성정보만을 이용한 경우에는 76%의 인식률을 나타내었으나 영상정보를 추가한 경우에는 88%의 인식률을 보여 12%의 인식률 증가가 있었음을 알 수 있다. 하지만 초성의 경우는 70% 정도의 상대적으로 낮은 인식률을 보이는데, 이것은 초성의 프레임수가 다른 음소에 비해 상대적으로 작아서 학습을 위한 데이터가 부족했던 것으로 보인다.

표 3은 평가용 데이터에 대한 숫자음별 인식결과이다.

| 데이터의 종류 | 음성정보 | | 음성&영상정보 | |
|---|---|---|---|---|
| | LPC | MFCC | LPC | MFCC |
| 공 | 2/5 | 3/5 | 3/5 | 4/5 |
| 일 | 3/5 | 3/5 | 3/5 | 5/5 |
| 이 | 5/5 | 4/5 | 4/5 | 5/5 |
| 삼 | 3/5 | 3/5 | 3/5 | 3/5 |
| 사 | 4/5 | 4/5 | 5/5 | 5/5 |
| 오 | 2/5 | 5/5 | 4/5 | 5/5 |
| 육 | 1/5 | 3/5 | 5/5 | 4/5 |
| 칠 | 2/5 | 2/5 | 1/5 | 1/5 |
| 팔 | 4/5 | 5/5 | 4/5 | 4/5 |
| 구 | 3/5 | 3/5 | 3/5 | 3/5 |

표 3 숫자음별 인식결과

숫자음별 인식결과를 보면 "삼"과 "칠"이 상대적으로 낮은 인식률을 보였는데, 오류를 분석하여 본 결과 "삼"은 "팔"로 "칠"은 "일"로 인식하는 경우가 많았다. 이것은 서로 중성과 종성이 비슷하거나 같은 구조로 구성되어 있기 때문으로 보인다. 그러므로 인식률을 높이기 위해서는 초성인식에 대한 개선이 필요 할 것이다.

## VI. 결 론

본 논문에서 구현한 인식시스템은 음성정보만을 이용한 시스템과 비교하여 보다나은 인식률을 보였다. 특히 학습된 화자의 경우에서는 100 %로 좋은 결과를 얻을 수 있었다. 또한 학습되지 않은 화자의 경우에도 평균 8%정도의 인식률 향상이 있었다.

따라서, 가시적인 정보가 보다 높은 인식률을 가지는 음성인식시스템의 설계에 있어서 상당히 유용한 입력 파라메타로 사용될 수 있음을 확인하였다.

이러한 연구를 바탕으로 향후 연속 음성인식 시스템으로 확장할 수 있을 것으로 기대된다.

## 참고문헌

[1] 박인정, 이천우, 남상엽, 김형배, "음성-영상정보의 통합처리에 의한 음성인식", 전자공학회지, pp.29-41, 1999.

[2] 오문식, "실시간 음성인식기 개발과 응용에 관한 연구", 석사학위논문, pp.6-8, 1998.

[3] 이충웅, "화자독립 격리단어 인식기의 개발에 관한 연구", 한국과학재단보고서, pp.10-32, 1989.

[4] 백준기, "첨단 영상 미디어 서비스와 영상복원기술", 대한전자공학회지, pp.28-39, 1996.

[5] 이황수, "신경회로 컴퓨터 이론·응용 및 구현", 한국과학기술원, 제5장 pp.1-8, 1988.