

---

# 음성과 영상 정보를 이용한 우리말 숫자 인식

이종혁\* · 최재원\*

Digit Recognition using Speech and Image Information

Jong-Hyeok Lee\* · Jae-Weon Choe\*

---

이 논문은 2000학년도 경성대학교 교비지원사업에 의하여 연구되었음.

---

## 요약

대부분 음성 인식 시스템에서는 음성 신호에서 추출한 특징 파라미터를 입력 정보로 하고 있다. 본 연구에서는 숫자 인식률을 높이기 위하여 음성 인식 시스템에 음성과 영상 정보를 동시에 이용할 수 있는 방법을 제안하였다.

실험을 통하여 음성정보만을 사용한 인식결과와 음성과 영상정보를 동시에 사용한 인식결과를 비교한 결과, 음성과 영상 정보를 동시에 입력했을 때 약 6%정도의 인식률의 증가를 가져옴을 알 수 있었다. 이를 통해 숫자 인식을 위해 음성정보만을 사용하는 것보다 영상정보를 같이 사용하는 것이 더욱 효과적임을 알 수 있었다.

## ABSTRACT

In the majority of case, speech recognition method tried recognition using only speech information. In order to highten the recognition rate, we proposed recognition system that recognige digit using speech and image information.

Through an experiment, this paper compared the recognition rate performed by existent speech recognition method and speech recognition method that includes image information. When we added the image information to the speech information, the speech recognition rate was increased about 6%.

This paper shows that adding image information to speech information is more effective than using only speech information in digit recognition.

## I. 서 론

음성은 사용의 편이성 및 자연성 면에서 다른 인터페이스에 비해 우수하여 여러 선진국에서는 지난 수십

년간 음성 인식에 관해 다양한 연구를 해오고 있다.

1960년대부터 음성의 발성에 대한 기초 연구가 수행되어 온 이래 기계에 의한 연속 음성 인식, 합성에는 아직 많은 과제가 남아있지만 최근 고립 단어 인식에

---

\* 경성대학교 전기전자 · 컴퓨터공학부

접수일자: 2002. 2. 18

있어서는 많은 발전이 있어 상용 제품도 등장하고 있다.[1]

기존에 발표한 문헌 중 정재선은 신경 회로망을 이용하여 단독 숫자음 인식에서 음성 특징 파라미터로는 12차의 LPC(Linear Prediction Coefficient)계수를 사용하였고 1명의 화자가 20회 발음하여 10회는 학습, 10회는 평가하여 평균 89.1% 인식률을 나타내었다.[2] 박인정 등은 음성 특징 파라미터로는 13차 LPC계수를, 영상 특징 파라미터로 256단계 히스토그램을 신경 회로망에 학습한 결과, 30dB의 잡음환경에서 모음의 98%의 인식률을 나타내었다.[3]

기존의 음성 인식 방법은 대부분 음성 자체만에 대한 특징 파라미터를 구하여 인식을 시도하였다. 이러한 방식은 입력 레벨, 배경 잡음, 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 보조해주어 청취자가 음향적인 음성을 이해하기 곤란할 때 유용한 정보 자원을 제공하여 준다.

본 논문에서는 종래에 이용되던 음성에서 얻어지는 특징 파라미터뿐만 아니라 음성을 발성할 시 얻을 수 있는 가시적 데이터에서 추출되는 파라미터와 함께 음성인식을 시도하였다. 실험 및 평가를 위한 데이터로는 단독 숫자음을 사용하며 음성 특징 파라미터는 12차 LPC 또는 MFCC(Mel Frequency Cepstrum Coefficients)계수로 하고 영상 특징 파라미터로 16단계 히스토그램을 사용하며 인식기는 신경 회로망을 이용하였다.

## II. 음성 정보 분석

음성의 대표적인 특징 중의 하나는 다양성이다. 이와 같은 성질은 음성의 의미정보 이외에도 화자의 음색, 감정 상태 등과 같은 정보도 포함하고 있기 때문이다.[4]

음성 분석이란 음성 데이터로부터 각 구간의 스펙트럼 특징을 잘 표현할 수 있는 특징 파라미터를 추출하는 것이다.

### 2.1 전처리 과정

전처리 과정(Pre-processing)은 환경 적응, 끝점 검출, 반향 제거, 잡음 제거와 같은 음성 신호를 본격적인 분석과정에 들어가기 전에 처리하는 과정이다. 이 중 끝점 검출은 입력된 디지털 음성 신호로부터 음성 인식에 필요한 음성 구간만을 검출하는 것으로 시작 및 끝점의 검출 성능은 인식 시스템의 최종 인식률에 직접적인 영향을 주게 되므로 신뢰성 있는 음성 인식 기의 구현에 반드시 선결되어야 하는 부분이다. 일반적으로 영교차율(Zero Crossing Rate) 값과 에너지(Energy) 값을 고려하여 수작업으로 끝점 검출을 하고 있다.[5]

영교차율은 발생 시 성대에 의해 음성 스펙트럼이 감쇄되므로 유성음의 에너지는 주로 3khz 이하에 집중되기 때문에 영교차율은 낮다. 또한, 북음에서는 주위 환경에 따라 약간 다르나 대체로 무성음보다는 큰 값을 갖는다. 그러므로 유·무성음의 구별에 사용될 수 있다.

단구간 에너지는 무성음에서는 작고 유성음에서는 크게 나타난다. 그러나 마찰음과 같은 일부 자음에서는 세기가 약한 모음보다 크게 나타날 수도 있다. 이에 따라 에너지는 유성음과 무성음의 구별에도 이용될 수 있으며, 높은 음질의 음성 신호에서 북음과 무성음의 구별에도 유용하게 쓰인다.

### 2.2 음성 특징 추출

특징추출은 끝점 검출된 구간의 음성 신호에서 가장 작은 용량으로 효과적으로 표현할 수 있는 특징 벡터를 추출하는 것이다. 일반적으로 특징 추출에 이용되는 방법으로는 선형 예측 분석에 의한 LPC 추출법과 켐스트럼 추출 방식에 의한 MFCC 추출법 등이 있다.[6]

LPC의 기본 개념은 시간  $n$ 에서 신호  $s_n$ 을 과거  $p$ 개의 신호로 균사화할 수 있다는데서 출발한다. 즉,

$$\widehat{s}_n = \sum_{k=1}^p a_k s_{n-k} \quad k=1, 2, 3, 4, 5, \dots, p \quad (1)$$

이 때, 원 신호,  $s(n)$ 과 균사화된 신호,  $\widehat{s}(n)$ 의 차이에 의한 에러,  $e(n)$ 은 다음과 같다.

$$\begin{aligned} e_n &= s_n - \hat{s}_n \\ &= \sum_{k=0}^K a_k s_{n-k} \end{aligned} \quad (2)$$

시간  $t_0$ 에서  $t_1$ 까지의 자승오차의 합,  $E$ 는

$$E = \sum_{n=t_0}^{t_1} e_n^2 \quad (3)$$

이고, 식(2)를 대입하여 정리하면, 식(3)은

$$E = \sum_{i=0}^K \sum_{k=0}^K a_i c_{ik} a_k \quad (4)$$

이 된다. 여기서

$$c_{ik} = \sum_{n=t_0}^{t_1} s_{n-i} s_{n-k} \quad (5)$$

$a$ 에 관하여 식(4)를 계산하기 위하여,  $a$ 로 편미분하고, 그 결과를 0으로 하여 구하게 된다.

$$\frac{\partial E}{\partial a_k} = 0, \quad k=1, 2, 3, \dots, p \quad (6)$$

캡스터럼 추출 방법 중 인간의 청각 인지과정을 고려한 방법이 MFCC추출법이다. Mel Frequency란 인간의 귀가 일정한 주파수 구간 내에서 그 사이의 여러 밴드가 합쳐진 소리는 동일 크기의 동일한 소리로 들린다는 것이다. 이를 반영하여 화자 변동에 무관한 견고한 음성 특징을 추출할 수 있다.[7]

인간의 청각특성을 고려하여 주파수축을 재변환한 것을 멜 척도(Mel Scale)라 하고, 주파수(f)와의 관계는 다음과 같이 표현된다.

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad (7)$$

멜 캡스터럼(Mel cepstrum)은 입력된 음성신호에 대해 FFT로 계산된 스펙트럼을 다시 멜 척도로 나누어진 필터뱅크를 사용하여 필터뱅크 출력을 구한 후 그 출력의 로그값에 DCT(Discrete Cosine Transform)를 적용하는 것이며 이때 유도된 계수가 MFCC가 된다.

다음 식(8)은 log를 적용하고 DCT를 수행하는 과정을 식으로 나타낸 것이다.

$$MFCC_m = \sqrt{\frac{2}{N}} \sum_{i=1}^N \left\{ \log(x(i)) \cos\left[\frac{2\pi m}{N}\left(i - \frac{1}{2}\right)\right] \right\} \quad (8)$$

$$0 \leq i < N \quad (N \text{은 Filter-bank Data})$$

$$1 \leq m \leq Q \quad (Q \text{는 MFCC 총 계수})$$

고차와 저차의 cepstrum의 값의 분포가 매우 큰 차이를 가지므로 cepstrum 계수들의 스케일을 조정해서 크기를 비슷하게 하기 위해 리프터링 과정을 거치며

이를 식으로 나타내면 다음과 같다.

$$Cm' = 1 + \frac{Q}{2} \sin\left[\frac{\pi * m}{Q}\right] Cm \quad (9)$$

$1 \leq m \leq Q$  ( $Q$ 는 MFCC 총 계수)

### III. 인식시스템의 구현

음성과 영상의 저장 시 추가되는 잡음이나 노이즈 등을 어느 정도 줄이고 또한 입력 데이터의 양을 줄임으로서 인식시스템의 속도를 향상시킬 수 있으므로[8] 저장된 음성과 영상 정보에서 그 특정 값들을 추출하고, 이러한 특징들은 초성, 중성, 종성 각 음소 별로 구분하여 초성, 중성, 종성의 각각의 신경회로망의 입력 패턴으로 사용하였다. 신경회로망은 입력된 특정 값을 학습, 평가하여 인식 결과를 나타내도록 하였다.

#### 3.1 음성특징 추출

음성의 경우 3.4Khz 정도의 대역폭만 가지면 알아듣는데 저장이 없으므로 음성특징파라미터를 추출하기 위하여 음성은 샘플링 주파수 8Khz에 16bits 양자화 하여 저장하였다. 이렇게 저장된 데이터를 중 실세 음성구간만을 추출하기 위하여 전처리 과정을 기쳤다.

전처리 과정에서는 에너지 값과 영교차율 값을 구한 후, 이를 고려하여 수작업으로 시작점과 끝점을 검출하여 묵음구간을 제거하였다.

전 처리과정을 거친 음성 정보에서 특정 파라미터를 추출하기 위하여 LPC와 MFCC 두 가지 방법을 사용하였다. 두 가지의 추출 방법들은 각각 12차의 특징 값이 나오도록 구성하고 실험에서 각 특징 파라미터의 성능을 비교하였다.

#### 3.2 영상특징 추출

음성 저장과 동시에 발성시의 입술 모양을 영상의 크기 256\*256, 초당 15프레임 속도의 Gray-Image로 저장한다. 이렇게 저장한 영상 정보를 히스토그램 평활화를 통하여 명암 값이 균일한 이미지로 변환한 후 캐니언자를 사용하여 잡음을 강한 유판선을 추출하고 히스토그램을 이용하여 256개의 수치 값으로 변환하였다.

256개의 데이터는 실제 신경회로망의 입력데이터로

사용하기에는 많은 양이므로 다시 16개의 데이터로 줄여 인식을 위한 입력데이터로 사용하였다. 그림 1은 영상 정보의 추출과정을 나타낸 것이다.

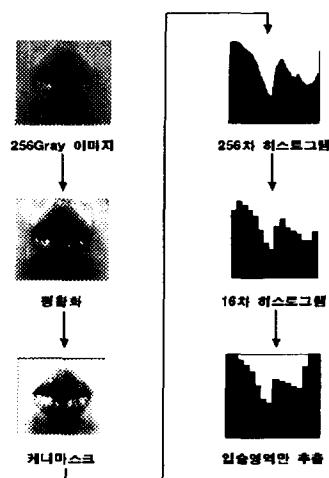


그림 1. 영상 정보의 추출과정

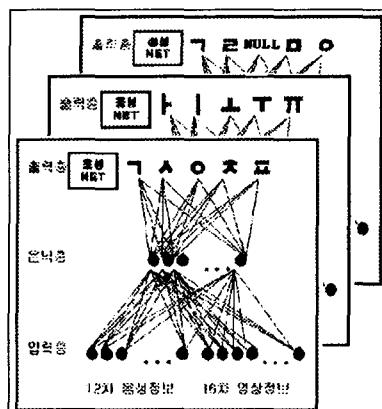


그림 2. 숫자음 인식 신경 회로망

### 3.3 신경회로망의 구성

인식기는 음성에서 구분 되어진 초성, 중성, 종성 데이터를 위하여 총 3개의 독립된 신경회로망으로 구성하였다. 초성, 중성, 종성 각각의 회로망에서 입력 층에는 30개의 노드를 두어 음성에서 12차의 LPC 또는 MFCC계수, 에너지와 영교차율, 그리고 영상에서 입술모양에 대한 16차의 히스토그램을 입력할 수 있도록

하였다. 은닉층에는 10개의 노드를 두어 회로망을 보다 견고하게 하였고, 출력층은 각 회로망마다 서로 다른 목표 출력 값을 갖는데, 우선 초성NET는 ㄱ, ㅅ, ㅇ, ㅊ, ㅍ 총 5개, 중성NET는 ㅏ, ㅓ, ㅜ, ㅠ, ㅣ 총 5개, 종성NET는 ㄱ, ㄹ, ㅁ, ㅇ, null로 구성하여 5개의 출력 값을 가지도록 회로망을 구성하였다. 그림 2는 실험에서 사용한 신경회로망의構成을 나타내었다.

## IV. 실험결과 및 검토

실험에 사용된 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. 기초실험을 통하여 오류 역전파 알고리즘에서의 학습률은 1, 모멘텀 변수는 0.1, 이득률은 0.05로 설정하였다. 반복 횟수는 실험 결과 5000번일 때 보다 2,500번 반복 학습시켰을 때 소요되는 시간도 적었고, 인식률의 차이도 0.1%정도의 차이로 아주 작았다. 따라서 반복 횟수는 2,500번으로 하였다.

그림 3은 학습화자 5명에 평가화자 5명을 랜덤하게 선택한 후 인식시스템을 통해 학습되어진 화자의 대한 인식률과 학습시키지 않은 평가용 화자의 인식률을 비교한 것이다.

학습된 화자의 인식 결과는 모든 경우에서 100%의 인식률을 보였다. 평가용 화자의 결과, 음성 정보만을 사용한 방법에서는 LPC를 이용한 경우에는 66%의 인식률을 나타내고, MFCC를 이용한 경우에는 76%의 인식률을 나타내었다. 영상 정보를 첨가한 방법에서는 LPC를 이용한 경우에 72%의 인식률, MFCC를 이용한 경우에는 82%의 인식률을 보여 영상을 추가한 인식방법이 6% 정도의 인식률을 증가가 있음을 알 수 있었다. 또한 MFCC를 방법이 LPC를 이용한 경우보다 평균 6% 높은 인식률을 보이고 있다.

학습은 20대 남성과 여성을 각각 3명과 2명으로 하고 평가용 데이터로는 남성과 여성을 각각 4명과 1명으로 한 후 MFCC를 사용하여 실험하였다. 학습 화자 5명에 대한 인식은 100%로 모두 인식되었고, 평가 화자의 경우, 초성, 중성, 종성의 각 음소별 인식 결과를 그림 4에 나타내었다.

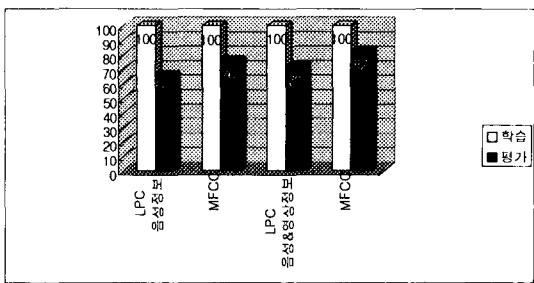


그림 3. 학습화자와 평가화자의 인식 결과

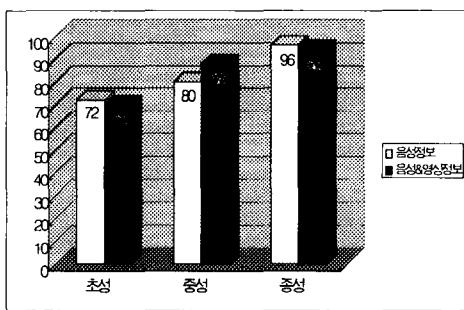


그림 4. 각 음소별 인식 결과

평가 화자에 대한 종성의 인식률을 살펴보면 94% 정도의 좋은 결과를 보이는 것을 알 수 있고, 종성의 경우는 음성 정보만을 이용한 경우에는 80%의 인식률을 나타내었으나 영상 정보를 추가한 경우에는 88%의 인식률을 보여 8%의 인식률 증가가 있었음을 알 수 있었다. 하지만 초성의 경우는 70%정도의 상대적으로 낮은 인식률을 보이는데, 이것은 초성의 프레임수가 다른 음소에 비해 상대적으로 작아서 학습을 위한 데 이터가 부족했던 것으로 보인다.

위 실험에서 평가용 데이터에 대한 숫자음별 인식 결과를 보면 표 1과 같다.

데이터의 종류별 인식 결과를 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 초성 인식에 대한 개선이 필요 할 것으로 생각된다.

표 1. 숫자음별 인식 결과

데이터의 종류	음성 정보		음성&영상 정보	
	LPC	MFCC	LPC	MFCC
공	3/5	3/5	3/5	4/5
일	3/5	4/5	3/5	5/5
이	5/5	4/5	4/5	5/5
삼	3/5	4/5	4/5	3/5
사	4/5	4/5	5/5	5/5
오	5/5	5/5	4/5	5/5
육	2/5	3/5	5/5	4/5
칠	2/5	3/5	1/5	2/5
팔	4/5	5/5	4/5	5/5
구	2/5	3/5	3/5	3/5

## V. 결론

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

특정 파라미터로 LPC계수, MFCC계수, 입술영역의 히스토그램을 사용하였으며 신경회로망을 사용하여 음성인식시스템을 구현한 후, 영상 정보의 효율성을 비교 검토했다.

실험 결과 MFCC 계수를 입력 패턴으로 사용한 것이 LPC 계수를 입력한 경우보다 10%정도의 더 좋은 인식률을 얻을 수 있었다. 또한, 음성만을 입력 패턴으로 사용한 경우보다 영상을 추가한 경우가 평균 6%의 인식률 향상이 있었다. 따라서 가시적인 정보가 보다 높은 인식률을 가지는 음성 인식 시스템의 설계에 있어서 상당히 유용한 특정 파라미터로 사용될 수 있음을 확인하였다.

그러나 저장된 입술 패턴의 기울어짐, 거리와 조명 등에 따른 영상 정보의 특징을 추출하는데 어려움이 많았다.

## 참고 문헌

- [1] 김수훈, “신경망 예측 HMM을 이용한 음성 인식에

- 관한 연구”, 동아대학교 박사학위논문, pp. 38-39, 1998.
- ※ 주관심분야 : 정보통신, 이동통신, 운영체제, 인터넷  
응용
- [2] 정재선, “신경회로망을 이용한 우리말 숫자음 인식”, 충남대학교 석사학위논문, 1999.
- [3] 박인정, 이천우, 남상엽, 김형배, “음성-영상 정보의 통합처리에 의한 음성 인식”, 전자공학회지, Vol. 26, No. 7, pp. 29-41, 1999.
- [4] 오문식, “실시간 음성 인식기 개발과 응용에 관한 연구”, 아주대학교 석사학위논문, pp. 6-8, 1998.
- [5] 최형기, “음성 인식을 위한 실시간 끝점 검출기의 구현”, 서울대 반도체 공학연구소, 연구결과보고서, pp. 7-10, 1998.
- [6] 박재형, “WWW상에서의 음성 처리에 관한 연구”, 건국대 석사학위논문, pp. 10-13, 1997.
- [7] 이충웅, “화자독립 격리단어 인식기의 개발에 관한 연구”, 한국과학재단보고서, pp. 10-32, 1989.
- [8] 박정훈, “자동차 내에서의 음성 인식에 관한 연구”, 전북대학교 석사학위논문, pp. 6-8, 2000.

## 저자소개

이종혁(Jong-Hyeok Lee)

1975년 2월 부산대학교 전자공학과(공학사)  
1980년 2월 부산대학교 대학원 전자공학과(공학석사)  
1991년 2월 부산대학교 대학원 전자공학과(공학박사)  
1980년 3월 ~ 1990년 2월 동의공업대학 전자과 부교수  
1990년 3월 ~ 현재 경성대학교 전기전자 · 컴퓨터공학  
과 교수  
1998년 7월 ~ 1999년 6월 미국 Beckman Institute,  
University of Illinois, 객원연구원

※ 주관심분야 : 인공지능, 음성인식, 신호처리

최재원(Jae-Weon Choe)

1988년 2월 고려대학교 컴퓨터공학과(공학사)  
1990년 8월 미시건주립대학교 컴퓨터공학과(공학석사)  
1995년 8월 건국대학교 전자공학과(공학박사)  
1990년 10월 ~ 1997년 8월 삼성전자 정보통신연구소  
선임연구원  
1997년 9월 ~ 현재 경성대학교 전기전자 · 컴퓨터공학  
과 조교수