論文2000-37TE-12-8

# 필터뱅크를 이용한 한국어 숫자음 인식 다이얼링 시스템

# (Korean Digit Speech Recognition Dialing System using Filter Bank)

朴 基 榮 *, 崔 炯 箕 **, 金 鍾 珖 **

(Kee-Young Park, Hyung-Ki Choi, and Chong-Kyo Kim)

요 약

본 논문은 한국어 숫자음 인식을 HMM과 DTW 프로그램을 사용한 필터 뱅크로 수행하였다. 스펙트럼 분석은 주로 성도의 모양에 의한 음성 신호 특징을 나타낸다. 그리고 음성의 스펙트럼 특징은 일반적으로 정의된 주파수 범위에서 적절하게 집중된 스펙트럼, 즉 필터뱅크를 통과해 나가는 것에 의해 얻을 수 있다. 또한 8개의 밴드 패스 필터는 인간 귀의 지각적인 청취력에 의해 나누었다. 정의된 주파수 범위는 320-330, 450-460, 640-650, 840-850, 900-1000, 1100-1200, 2000-2100, 3900-4000Hz이고, 샘플링 주파수는 8KHz이다. 그리고 프레임 폭은 20ms, 주기는 10ms이다. 실험 결과는 한국어 숫자음 음성인식에 대해 필터 뱅크를 사용하는 경우 HMM보다 DTW의 인식율이 더 높은 인식율이 나오는 것을 확인 할 수가 있었다. 필터 뱅크를 이용한 한국어 숫자음 인식율은 24차 밴드패스필터에서 93.3%, 16차 밴드패스필터에서, 89.1%, 8차 밴드패스필터의 하드웨어 음성 다이얼링 시스템에서 88.9%의 인식율을 나타내었다.

**Abstract**

In this study, speech recognition for Korean digit is performed using filter bank which is programmed discrete HMM and DTW. Spectral analysis reveals speech signal features which are mainly due to the shape of the vocal tract. And spectral feature of speech are generally obtained as the exit of filter banks, which properly integrated a spectrum at defined frequency ranges. A set of 8 band pass filters is generally used since it simulates human ear processing. And defined frequency ranges are 320-330, 450-460, 640-650, 840-850, 900-1000, 1100-1200, 2000-2100, 3900-4000Hz and then sampled at 8KHz of sampling rate. Frame width is 20ms and period is 10ms. Accordingly, we found that the recognition rate of DTW is better than HMM for Korean digit speech in the experimental result. Recognition accuracy of Korean digit speech using filter bank is 93.3% for the 24th BPF, 89.1% for the 16th BPF and 88.9% for the 8th BPF of hardware realization of voice dialing system.

## Ⅰ. 서 론

최근 음성과 자연언어의 기본적인 성질의 이해에 관한 관심이 높아지고 있고, 각종 미디어의 발달과 초고속 정보 통신망의 구축으로 통신을 통한 통신판매, 물류 처리 및 제품 홍보 등이 폭증하고 있다. 이와 더불어 컴퓨터의 보급에 의한 신호처리 기술과 정보처리 기술의 급속한 발전과 더불어 음성을 통한 인간과 기계사이에서의 정보 정합의 중요성이 증대되기 때문에 음성인식 시스템의 개발이 매우 중요하다.

그러므로 음성인식의 한 분야인 음성입력에 의한 숫자인식은 음성 다이얼링 서비스나 시각장애자를 위한 승강기 서비스, 신용카드 번호입력, 자동차 내에서의 다

* 正會員, 全州工業大學 情報通信科

(Dept. of Information & Communication Engineering, Jeonju Technical College)

** 正會員, 全北大學校 電子工學科

(Dept. of Electronic Eng. Chonbuk National Univ.)

接受日字:2000年10月24日, 수정완료일:2000年11月30日

이얼링 등과 같은 실생활의 여러 부분에서 활용 될 수 있다.

R. Bellman이 1957년 DP(Dynamic Programming)라는 수학적 기법을 제시한 후 DP기술에 관한 많은 연구가 진행되고 있다.[1][2] 특히 음성인식에서 화자의 speaking rate변화로 인한 시간축 상에서 음성 패턴의 비선형 변동을 어떻게 시간 정규화시켜 인식율을 개선하느냐 하는 문제는 매우 중요하다.[3][4][5]

이러한 음성기술의 발달에 힘입어 미국에서는 1971년부터, 일본은 1982년 그리고 유럽 국가들은 1984년부터 소프트웨어(DTW, HMM, NN) 등으로 화자종속 고립단어, 연속음성 인식율이 90~98%까지 이르고 있다.[2] 그러나 국내에서는 1992년이후 경북대, 목포대, 전북대, ETRI에서 숫자음 인식을 LPC켑스트럼과 HMM으로 처리하여 72~93.78%의 인식율을 나타내고 있지만 하드웨어로 구성한 실시간 인식시스템의 실용화 단계에 이르기에는 크게 미흡한 상태에 있다.

더욱이 전화는 현대 생활에 없어서는 안될 필수품으로써, 최근에는 단순 음성 전달목적을 넘어 위성 통신을 통한 다양한 형태의 정보를 전송할 수 있는 초고속 통신망을 구축하기 위한 연구가 활발하게 진행되고 있다. 그러나 전화를 걸 때, 과거의 기계식 다이얼링 접점식에서 현재의 복합 주파수(MFC)다이얼링으로밖에 발전하지 못했다. 아울러 음성인식을 위한 하드웨어 설계에서의 중요한 관점은 실시간 처리의 신속성, 편리한 이용을 위한 기능성 그리고 범용화를 통한 경제성 등에 초점을 두고 있다. 이러한 현실적 문제의 구현을 위해서는 여러 가지의 알고리즘 등을 통한 프로그램의 개발과 하드웨어 구성으로 이루어져야 할 것이다.[6]

본 논문에서는 음성 다이얼링을 위한 화자종속 한국어 숫자음 인식을 위해 지금까지 프로그램을 통한 전처리로 인해 실시간 처리에 많은 지장을 초래했던 과정을 필터뱅크를 이용한 하드웨어 시스템을 구축하여 실시간화 한다. 즉 음성 다이얼링 시스템에서 통화하고자 하는 상대방의 전화번호를 누르는 대신에 음성(화자종속)으로 전화번호(고립숫자)를 입력하면 음성다이얼링 시스템에서 필터링, buffer & line driver회로, 래치회로, 멀티플렉서회로와 개인용 컴퓨터의 인식프로그램(DTW알고리즘을 사용하여 DP 인식)을 통하여 전자다이얼링이 되어 상대방과 쉽게 통화할 수 있는 자동 음성 다이얼링 시스템을 구성하는 것을 의미한다. 그리고 음성 다이얼링 시스템을 통과한 음성은 비트 레벨에 의한 DTW알고리즘의 정수 연산으로 특징 추출시 많은 계산량을 줄인다. 또한 확률계산식의 계산량 감소 방법을 통해 인식시간을 줄이고 이러한 방법들이 실시간 인식시스템의 구현에 효과적임을 보인다.

## Ⅱ. 한국어 숫자음성 특징추출

### 1. 필터 뱅크(Filter Bank)

음성 인식에 대해 신호처리 프론트 엔드(signal processing front end)에는 2가지 방식이 있는데, 하나는 그림 1의 필터뱅크 분석모델이고, 다른 하나는 그림 2의 LPC 분석모델이다. 이 모델들은 실질적인 음성인식 시스템에 있어서 고효율을 나타내고 있음을 증명하고 있다.[7] 그리고 그림에서 신호 $s(n)$은 Q개의 밴드패스필터(band pass filter : BPF)를 통과한 이산 신호를 나타내고 있으며, 본 논문에서 Q개의 신호에 대한 주파수의 범위는 320-330, 450-460, 640-650, 840-850, 900-1000, 1100-1200, 2000-2100, 3900-4000Hz로 설정하였다. 또한, 각각의 필터 뱅크는 그림 1의 아래와 같이 주파수의 범위를 서로 중복시킨다.

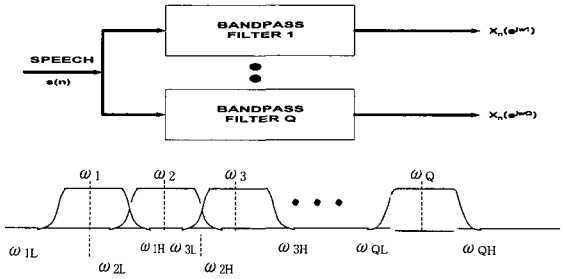


그림 1. 필터뱅크 분석모델
Fig. 1. Bank-of-filters analysis model.

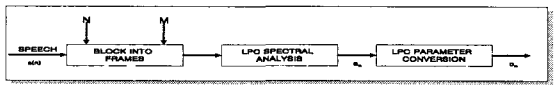


그림 2. LPC 분석모델
Fig. 2. LPC Analysis model

이때 BPF의 $i$번째의 출력은 $X_n(e^{jw_i})$으로서, 신호 $s(n)$의 단구간 스펙트럼에 관한 표현식으로 나타난다. 여기서 $w_i$는 정규화 주파수 $2\pi f_i / F_s$이고, $F_s$는 샘플링 주파수이다. 즉 임의의 시간 $n$에서 중심주파수 $w_i$인 BPF의 $i$번째 출력임을 알 수 있다.

완전한 필터 뱅크 프론트 엔드 분석기에서 Q개의 BPF를 통과한 뱅크의 표본화 음성 신호를 $s(n)$, 표본화 수 $M_i$, BPF의 $i$번째의 임펄스 출력 $h_i$ BPF를 통과한 음성 출력 $S_i(n)$은

$$S_i(n) = S(n) * h_i(n), \quad 1 \le i \le Q$$
$$= \sum_{m=0}^{M_i-1} h_i(m)s(n-m) \tag{1}$$

이다. 위 식에서 나타내는 의미는 각각의 BPF 신호 $S_i(n)$은 주어진 필터 주파수에서의 음성 신호 에너지를 측정한다는 사실을 알 수 있다.[7]

2. 특징파라메타

이 절에서는 숫자에 대한 특징파라미터에 대해서 설명하기로 한다. 현재의 프레임에 대해서 숫자의 특징을 잘 나타내는 파라미터를 선정하는 문제는 인식시에 인식률과 직결되는 사항이다.[1,2] 그러므로 특징파라미터를 추출하기 위해 그림 3에 나타낸 방법으로 먼저 시뮬레이션(simulation)을 한 다음, 이것을 토대로 하여 특징이 나타난 주파수 대역의 특징을 필터 뱅크화 하여 숫자음의 특징을 추출하여 인식하도록 하였다. 먼저 시뮬레이션을 통한 숫자음의 특징추출 방법은 다음과 같다. 즉 한국어 숫자 음성 신호를 전역필터링 하고, 여기에 음성신호의 특징을 가장 잘 추출 할 수 있는 해밍 창(hamming window)함수를 가한다. 이때 프레임의 폭은 20ms이며 프레임의 주기는 10ms를 이용하고 있다. 이러한 조건을 이용하는 경우 실시간으로 동작하는 인식기를 구현하였을 때 정확하고, 확실한 인식을 위한 8차 특징 하나를 얻기 위해서 주파수 필터 뱅크를 인간 귀의 지각적인(perceptual) 청취력에 의해 1000Hz 이하에서 5단계로(320-330/450-460/640-650/840-850/900-1000Hz) 세분화하고, 1000Hz 이상에서는 3단계(1100-1200/2000-2100/3900-4000Hz)로 넓게 구성하여, 필터 뱅크의 8차특징(시뮬레이션에서는 16차, 24차를 사용함)의 8KHz 표본화된 특징을 추출하였다. 그리고 FFT(fast fourier transform)와 viterbi 알고리즘을 통한 인식과 DTW와 DP를 이용한 인식을 시행하였다. 또한 필터뱅크를 통한 이산 값이 본 논문에서 제시한 하드웨어 시스템을 통과하면서 정수 값으로 계산되므로 인식 알고리즘을 사용 할 때 underflow가 발생하지를 않아 실시간 인식의 가능성을 제시 해 준다. 한편, 특징 차수의 증가는 인식률과 항시 비례하지 않으며 오히려 실시간 인식시에 계산량의 증가로 인한 인식시간의 지연이 발생할 수 있으므로 실시간 인식에서는 적정한 차수로 결정하는 것이 중요하다.[8]

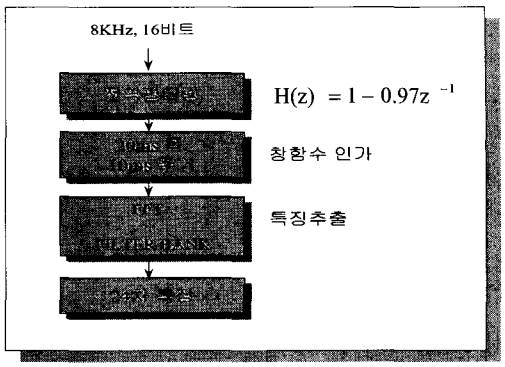


그림 3. 전체 특징추출의 블록도
Fig. 3. Feature extraction blocks.

## Ⅲ. 실험 방법 및 결과

1. 하드웨어 구성

이 논문에서는 20-40세가량의 남성화자 20명의 음성을 잡음이 있는 실험실에서 자연스럽게 발음한 한국어 숫자음 /일/, /이/, /삼/, /사/, /오/, /육/, /칠/, /팔/, /구/, /영/ 10개를 각각 4회씩 발성한 총 800개의 음성 데이터로 사용한다. 음성 인식을 위한 전체 블럭다이어그램은 그림 4와 같다.

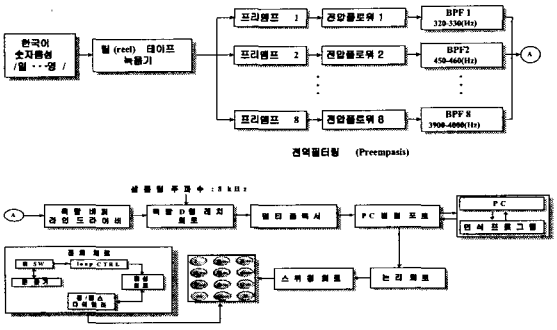


그림 4. 인식 블럭 다이어그램
Fig. 4. A block diagram of recognition.

한국어 숫자음을 콘덴서 마이크로폰(condenser microphone EM-550)을 사용하여 주파수 특성이 20(Hz)에서 20(KHz)까지 선형성을 지닌 성능이 우수한

릴 테이프 녹음기(reel tape recoder : MODEL NO. TEAC X-300R)에 녹음하였다. 이때 녹음된 장소는 잡음이 있는 실험실에서 자연스럽게 녹음을 하였다. 녹음기에서 나온 음성 신호는 최대 출력 전압이 ±1(V)가 되도록 조절하여 하드웨어로 구성한 8채널의 프리앰프(4.6KHz의 저역여파기와 선형증폭기)와 전압플로워를 통과한 후 필터뱅크/320-330Hz, 450-460Hz, 640- 650Hz, 840-850Hz, 900-1000Hz, 1100-1200Hz, 2000- 2100Hz, 3900-4000Hz/를 통과시킨다. 여기서 나온 아날로그 숫자음 신호를 연산증폭기로 만들어진 8개의 밴드로 구성된 필터뱅크를 통과하고, 이 아날로그 신호를 다시 8채널의 옥탈버퍼 라인드라이버(octal buffers & line driver with 3-state output)를 통과시킨다. 그러면 이곳에서 출력된 숫자 음성 신호는 이산 신호가 된다. 이 신호를 상한 주파수가 8KHz로 샘플링 되도록 클럭 주파수를 입력시킨 옥탈 D형 래치 회로(octal D type transparent latch & edge-triggered FFs)를 통과한다. 그 결과 8bit로 양자화한 특징 데이터가 생성된다. 그리고 이 데이터를 멀티플렉서(quadruple 2 line to 1 line data selectors/multiplexers)에 저장한다. 여기서 추출된 특징 데이터를 퍼스날 컴퓨터의 병렬포트(parallel port)를 통하여 PC의 포트에 저장된다. 포트에 저장된 특징 데이터를 인식 프로그램을 통해서 인식된 결과를 다시 퍼스날 컴퓨터의 병렬포트를 통하여 논리회로에 연결되고 여기서 다시 스위칭 회로에 연결되어 MFC(multi frequency control) 다이얼 패드(dial pad)의 C라인(coloumn line)과 R라인(row line)으로 인식된 숫자음의 선택점(contact point)이 연결되어 전화기의 DTMF(dual tone multi frequency)레벨과 출력 레벨을 인식하여 전화기의 자동 다이얼링이 이루어진다.[6]

2. 숫자음 인식프로그램

이 논문에서 사용한 인식 알고리즘은 2가지를 사용하였는데, 하나는 HMM을 이용한 Modified k-Mean 알고리즘을 사용하였고 다른 하나는 DTW와 DP 프로그램을 이용한 알고리즘을 이용하였다. Modified k-Mean을 사용한 알고리즘에서 M을 32와 64로 하여 코드북을 생성하였다. 이렇게 만들어진 코드북을 이용하여 모든 음성 데이터에 대해 상태 개수를 3으로 하고 Baum-Welch 알고리즘을 이용하여 상태 천이 확율 $a_{ij}$와 상태 $j$에서 입력 데이터가 코드북에서 발생 할 확율 $b_j(k)$를 구한다. 이 알고리즘은 Hidden Markov Model(HMM)[10]과 코드북의 작성[11]을 기반으로 하는 알고리즘으로서 그 구조는 그림 5와 같다

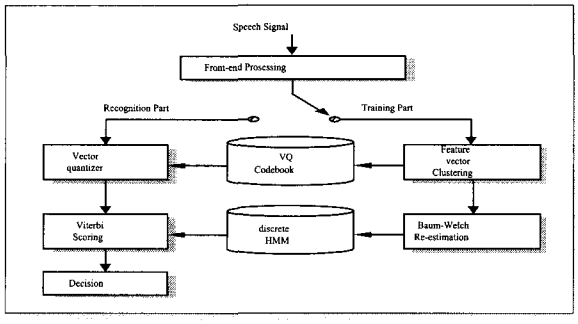


그림 5. 인식프로그램의 플로우 챠트
Fig. 5. Flowchart of Recognition Program.

한편, DTW와 DP 프로그래밍을 이용한 알고리즘을 이용한 프로그래밍은 비선형적인 시간축의 변동 패턴을 선형적으로 정규화 시키기 위해서 패턴 정합 방식을 사용한다. 이 알고리즘에서 시간축 변동은 몇가지 특정한 비선형성 워핑(warping)함수로 이루어져 있고 또 비교되는 두 음성의 패턴사이의 시간적 차이는 워핑함수를 이용하여 표준 패턴을 이 알고리즘에 적용하여 분석된 특징벡터의 최대 일치점에 도달하도록 표준 패턴의 시간축을 변화시켜 해결하였다. 이때의 정규화 거리는 두 패턴들 사이의 오차 거리로부터 계산하였고, 이 계산 거리의 최소화 처리는 DP프로그래밍을 사용하여 매우 효율적으로 수행하였다. 이중에서 학습단어는 각 화자가 1회 발성한 것을 사용하였으며, 학습과정에서 사용하는 DTW는 대칭 형태인 경로를 이용하였고

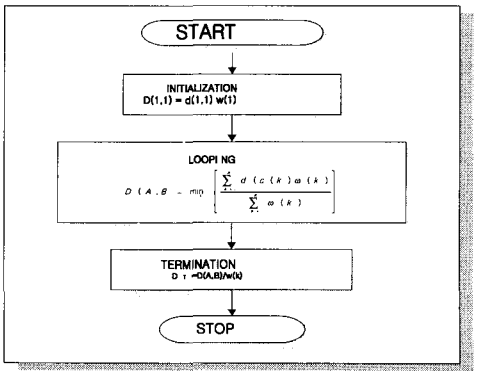


그림 6. DTW 알고리즘의 흐름도
Fig. 6. Flowchart of DTW algorithm.

정합창으로 하여금 기울기를 갖도록 하여 단어의 발성 길이의 차가 심한 경우 생기는 경로상의 오차를 줄이도록 하였다. 그리고 2회분 발성한 것으로는 다수 표준 패턴으로 이용하기 위해 세 개의 패턴을 작성하였으며, 나머지 1회 발성한 것을 시험패턴으로 하였다. DTW알고리즘의 흐름도는 그림 6과 같다

3. 결과 및 분석

먼저 필터 뱅크를 통과한 각 숫자음에 대한 주파수 특징을 그림 7에 나타내었다. 필터 뱅크를 통한 한국어 숫자 음성 신호의 특징을 추출하기 위하여 먼저 컴퓨터 프로그램을 통한 시뮬레이션을 실시하고, 그 결과를 기준으로 하여 하드웨어에 적용하여 비교 검토하였다. 컴퓨터 시뮬레이션은 한국어 숫자 음성 신호에 대해 고주파 성분의 감쇄를 보상할 목적으로 HPF(high pass filter)를 가하고, $\alpha$값을 0.98로 취하여 전역필터링을 수행하였다. 그리고 음성신호에 대해 동일한 구간 폭을 지정하여 그 구간에 해당하는 숫자음에 대해서만 음성의 특징을 추출한다. 음성 구간의 정확한 스펙트럼의 특징을 얻기 위해서 가장자리에서 작은 값을 갖고 중앙 부분에서 1에 가깝고, 음성신호의 특징을 가장 잘 추출 할 수 있는 해밍 창(hamming window)함수를 가했다. 이때 프레임의 폭은 20ms(160샘플)이며 프레임의 주기는 10ms(80샘플)로 하였다. 이러한 조건을 이용하는 경우 실시간으로 동작하는 인식기를 구현하였을 때 정확하고, 확실한 인식을 위한 16차, 24차로 시뮬레이션을 실험하였고 여기서 얻어진 특징을 바탕으로 하여, 하드웨어를 구성하였다. 여기서 필터의 구성은 24차의 시뮬레이션을 통한 필터 중 음성 특징을 가장 잘 나타내는 부분(표 1)을 기준으로 하여 다음과 같은 8개의 필터부분을 선택하여 구성하였다. 그리고 음성 특징을 얻기 위해서 주파수 필터 뱅크를 인간 귀의 지각적인(perceptual) 청취력에 의해 1000Hz 이하에서 5단계로 세분화 하였다. 그 필터 대역은 (320-330/450-460/640-650/840-850/900-1000Hz)세분화하고, 1000Hz 이상에서는 3단계(1100-1200/2000-2100/3900-4000Hz)로 넓게 구성하였다. 그리고 24차의 필터 뱅크 시뮬레이션에 사용한 각 숫자음의 필터 주파수 특징이 그림 7에 나타나 있다.

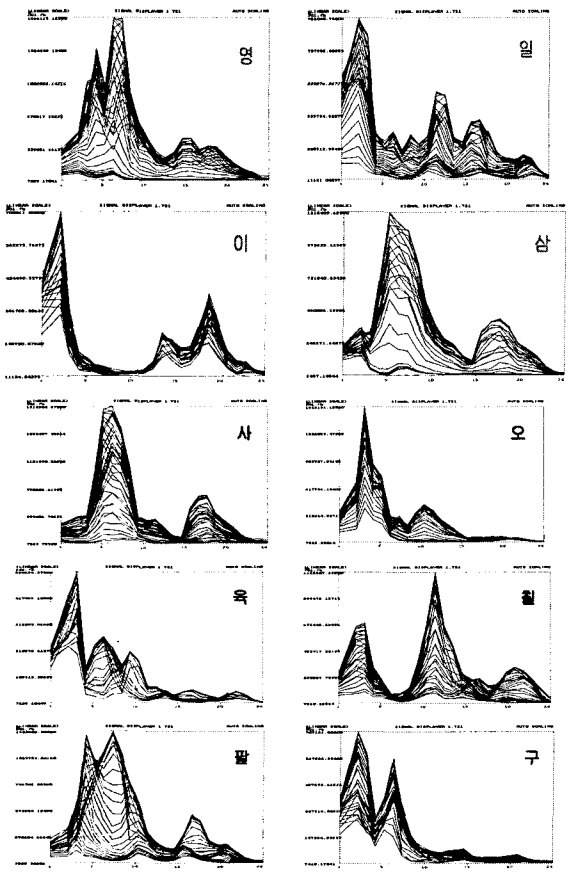


그림 7. 한국어 숫자음의 주파수 특징 추출
Fig. 7. Feature extraction of frequency for Korean digit speech.

표 1. 한국어 숫자음의 주파수 특성
Table 1. Frequency characteristics of korean digit speech (unit : Hz)

| | magnitude th. = 0.3 | magnitude th. = 0.4 |
|---|---|---|
| /영/ | 1000~1200 | 1000~1200 |
| /일/ | 350~450/1900~2100/2800~3000 | 350~450/1900~2100 |
| /이/ | 200~330/3000~3200 | 200~330/3000~3200 |
| /삼/ | 900~1500/2800~3000 | 900~1500 |
| /사/ | 900~1200/2600~2900 | 900~1200 |
| /오/ | 600~800/1600~1800 | 600~800 |
| /육/ | 400~700/900~1100/1600~1700 | 400~700/900~1100 |
| /칠/ | 400~600/1900~2100/3500 | 400~600/1900~2100 |
| /팔/ | 600~1400/1800~2000 | 600~1400 |
| /구/ | 350~500/900~1200 | 350~500/900~1200 |

주파수 특성상에서 중복되는 이유는 음향학적 측면에서 해석 해보면 다음과 같다. 숫자음 /일/, /이/, /칠/의 하위 주파수 특성이 약 500Hz 이하에서 중복되는

이유는 혀의 위치와 구강에 의해 결정된다. 그러므로 /일, 이, 칠/은 전설모음 '이'를 포함하고 있기 때문에 같은 주파수로 중복된다. 특히 /일/과 /칠/은 상위 주파수는 1900~2100Hz로 더욱 비슷한데 이것은 자음의 종성 부분에 유음 'ㄹ'을 포함 한 이유 때문인 것으로 생각된다. 또한 /삼/, /사/, /팔/의 하위 주파수가 600~1400Hz인 이유는 같은 맥락에서 중설모음 '아'를 포함한 이유라고 생각된다. 게다가 /삼/과/사/의 상위주파수 900~1200Hz로 더욱 많이 중복된 이유 또한, 자음의 마찰음 'ㅅ'을 포함하기 때문인 것으로 생각된다. 그리고 /오/와 /구/의 하위 주파수 350~500Hz로 중복되는 이유는 후설 모음 '오'와 '우'를 포함하기 때문이다. 마지막으로 숫자음 중에서 유별나게 음성 구간이 가장작고 쉽게 구분되는 /육/은 반모음 '유'음과 자음 중에서 파열음 'ㄱ'을 포함하기 때문이다. 이상의 필터 뱅크를 통과한 숫자음의 주파수 특징을 추출하기 위한 기준은 음성의 크기(magnitude)에 대한 문턱값(threshold value)를 얼마나 정확히 결정하느냐가 숫자음 인식의 중요한 변수로 작용함을 알 수 가 있다. 표 2와 표 3에서 보는 바와 같이 문턱값과 필터의 차수의 변화에 따라 인식율이 다르게 변화하는 것을 알 수 있다.

또한, 그림 8에서 나타난 바와 같이 필터 뱅크의 차수가 낮아도 문턱값과 필터 주파수를 정확히 설정해 줄 경우 인식율의 변화가 크게 나타나지 않음을 확인 할 수가 있다. 이런 결과로 인해 최소의 8차 필터뱅크를 구성하여 실험해 본 결과 표 4에서는 DTW를 이용한 문턱값과 인식율이 88.8%였으며, 표 5에서 나타난 HMM을 이용한 문턱값과 인식율은 85.8%였다. DTW와 HMM의 인식율에 대한 비교 (그림 9)에서 알 수 있는 바와 같이 숫자음의 인식율은 HMM을 사용하는 것보다 DTW를 사용하는 것이 시간적으로나, 인식율 면에서 월등하게 나타남을 확인 할 수가 있었다.

HMM에서 인식율이 나쁘게 된 이유는, 입력 음성에 대해서 하나의 특징을 이용하여 코드북을 만들어 인식하여야 한다. 그런데, 필터 뱅크를 통해서 출력된 이산 데이터는 제한적이므로 천이가 이루어 졌을 때 관측 대상으로부터 각 출력 심볼이 관측될 확률을 구하지 못하는 것을 실험을 통해 확인 할 수가 있었다. 그러므로 HMM의 가장 큰 장점인 확률적 접근 방식을 적용하여 만든 코드북으로 숫자 음성 신호를 확률적 모델링을 통한 인식이 되어야 하는데 입력 심볼과 양자화된 심볼간의 거리가 코드북을 특징 지우는 거리 값의 출력확률 밀도함수를 구하기기 어려워서 인식율의 저하를 갖는다고 생각된다. 이런 이유로 필터뱅크를 이용한 HMM인식은 적합하지 않다고 생각된다.

그림 10은 실제로 구성하고 제작한 8차 필터 뱅크와 실험에 사용한 실험장치를 나타내었으며, 8챈널 필터 뱅크를 구성한 하드웨어 시스템에서의 한국어 숫자음에 대한 인식율은 88.9%를 나타내었다. 그리고, 16챈널 필터 뱅크를 구성한 시뮬레이션에서는 89.1%, 24챈널 필터 뱅크를 구성한 시뮬레이션에서는 93.3%의 인식율을 얻었다.

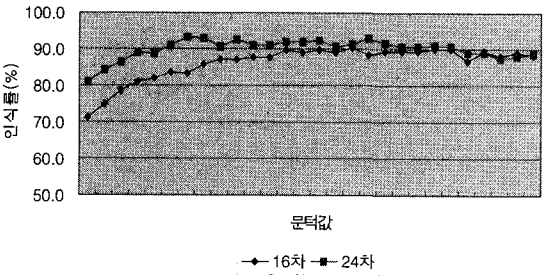


그림 8. 필터 뱅크의 문턱값과 인식율의 비교
Fig. 8. Comparison of recognition rate and threshold value of filter bank.

표 2. 16차 필터 뱅크의 문턱값과 인식율의 비교
Table 2. Comparison of recognition rate and threshold value in 16th order filter bank.

| threshold value | recognition rate |
|---|---|
| 10000.000000 | 15.555555% |
| 15000.000000 | 32.222221% |
| 20000.000000 | 63.333332% |
| 30000.000000 | 74.166664% |
| 40000.000000 | 79.722221% |
| 50000.000000 | 81.388885% |
| 60000.000000 | 86.111115% |
| 75000.000000 | 87.777779% |
| 85000.000000 | 89.166664% |
| 95000.000000 | 87.500000% |
| 100000.000000 | 88.611115% |
| 105000.000000 | 89.166664% |
| 110000.000000 | 89.722221% |
| 115000.000000 | 89.166664% |
| 125000.000000 | 91.111115% |

표 3. 24차 필터 뱅크의 문턱값과 인식율의 비교

Table 3. Comparison of recognition rate and threshold value in 24th order filter bank.

| threshold value | recognition rate |
|---|---|
| 10000.000000 | 40.555557% |
| 15000.000000 | 65.000000% |
| 20000.000000 | 74.166664% |
| 30000.000000 | 83.611115% |
| 40000.000000 | 90.000000% |
| 50000.000000 | 90.555557% |
| 55000.000000 | 93.333336% |
| 65000.000000 | 92.222221% |
| 70000.000000 | 92.777779% |
| 95000.000000 | 92.222221% |
| 100000.000000 | 92.222221% |
| 105000.000000 | 91.944443% |
| 110000.000000 | 92.222221% |
| 115000.000000 | 90.833336% |
| 125000.000000 | 91.111115% |

표 4. 8차 필터 뱅크의 문턱값과 인식율의 비교

Table 4. Comparison of recognition rate and threshold value in 8th order filter bank.

| threshold value | recognition rate |
|---|---|
| 10000.000000 | 82.777779% |
| 110000.000000 | 84.722221% |
| 120000.000000 | 84.444443% |
| 125000.000000 | 84.444443% |
| 130000.000000 | 84.722221% |
| 134000.000000 | 85.277779% |
| 135000.000000 | 85.555557% |
| 136000.000000 | 86.388885% |
| 137000.000000 | 88.333336% |
| 138000.000000 | 88.888885% |
| 139000.000000 | 86.944443% |
| 140000.000000 | 85.833336% |
| 147000.000000 | 85.833336% |
| 150000.000000 | 86.111115% |
| 57000.000000 | 85.555557% |

표 5. 8차 HMM의 문턱값과 인식율의 비교

Table 5. Comparison of recognition rate and threshold value in 8th order HMM.

| threshold value | recognition rate |
|---|---|
| 100000.000000 | 77.160496% |
| 110000.000000 | 80.246913% |
| 120000.000000 | 83.641976% |
| 130000.000000 | 82.098764% |
| 140000.000000 | 83.333331% |
| 146000.000000 | 84.567899% |
| 147000.000000 | 85.802472% |
| 148000.000000 | 85.493827% |
| 149000.000000 | 84.259260% |
| 150000.000000 | 84.259260% |
| 151000.000000 | 84.259260% |
| 160000.000000 | 82.716048% |
| 170000.000000 | 82.716048% |
| 180000.000000 | 82.098764% |
| 182000.000000 | 78.395063% |

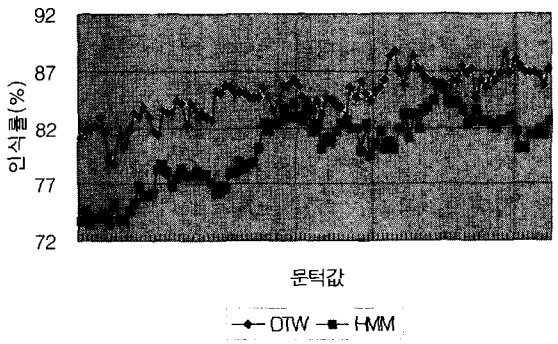


그림 9. DTW와 HMM 인식율의 비교

Fig. 9. Comparison of DTW and HMM recognition rate.

그림 10. 음성다이얼링 시스템 사진

Fig. 10. Picture of voice dialing system.

## Ⅳ. 결 론

이상에서 실험한 결과와 분석을 통해 본 논문에서 제시한 하드웨어의 구성과 설계, 그리고 프로그래밍을 통한 인식에 대한 결론은 다음과 같다. 첫째로 인간 귀의 지각적인 주파수 대역폭으로 나눈 8채널의 필터 뱅크 회로, 옥탈버퍼 라인 드라이버, 옥탈 D형 래치 회로, 멀티플렉서, 논리회로와 스위칭회로를 구성하여 설계하였다. 하드웨어 설계와 구성에서 가장 큰 잇점은 시스템을 통과한 이산 데이터 값이 모두 정수이므로 인식 알고리즘을 이용할 때 underflow가 발생하지 않아 실시간 인식에 커다란 전환점을 제시 해 준다. 둘째로는 HMM과 코드북의 작성을 기본으로 한 알고리즘과 DTW와 DP알고리즘을 통한 프로그래밍을 통한 인식율에 대한 비교(그림 9)에서 알 수 있는 바와 같이 숫자음의 인식율은 HMM을 사용하는 것보다 DTW를 사용하는 것이 시간적으로나, 인식율면에서 많은 차이가 나는 것을 확인 할 수 있었다. 이런 이유로 필터뱅크를 이용한 HMM인식은 적합하지 않다고 결론지었다. 한편으로는 DTW와 DP알고리즘을 통한 인식 프로그래밍에서 필터 뱅크를 통한 평행사변형 꼴의 전체 경로 제약을 DTW알고리즘에 적용시킴으로써 모든 소구간 경로 제약의 알고리즘에 대한 인식 수행 시간을 단축 시킬 수 있다. 그리고 시뮬레이션의 결과를 통해 알 수 있듯이 필터 뱅크의 차수의 증가는 인식율과 항상 비례하지 않으며, 오히려 실시간 인식시에 계산량의 증가로 인한 인식 시간의 지연이 발생 할 수 있으므로 실시간 인식에서는 적절한 차수로 결정하는 것이 중요함을 확인 할 수 가 있었다. 결론적으로 본 최종 인식율은 8채널 필터 뱅크를 구성한 하드웨어 시스템에서는 88.9%, 16채널 필터 뱅크를 구성한 시뮬레이션에서도 89.1%, 24채널 필터 뱅크를 구성한 시뮬레이션에서는 93.3%의 인식율을 얻었다.

이 논문에서 얻어진 결과로 미루어 볼 때 한국어 숫자음 인식에 대해서는 보다 정확한 필터뱅크의 하드웨어 구성과 OP앰프의 오프셋 전압에 대한 문제점 제거가 선행되어야 할 것이다. 또한 프로그래밍에서도 Neural Network, Continuous HMM 또는 복합적인 프로그래밍(NN + HMM)을 통한 소프트웨어적인 실험이 추후 연구되어야 할 것이다. 그리고 완전한 하드웨어 시스템을 구성하기 위해서는 더 신뢰성 높은 최적 DP 알고리즘을 구하기 위해 더 많은 표준 패턴을 설정하고, 모든 제한 조건 및 여러 가지 거리 계산 방법을 적용 분석하여야 할 필요가 있으며, 더불어 실시간 처리 및 완벽한 인식율을 위한 새로운 DP알고리즘으로 MPW(multi-project wafer)화 하여 하드웨어를 구성하여야 할 것이다. 결국 완전한 하드웨어 구성이 이루어지면, 한국어 숫자음 음성입력에 의한 숫자인식은 음성 다이얼링 서비스나 시각장애자를 위한 승강기 서비스, 신용카드 번호입력, 음성 엘리베이터, 음성 우편분리기, 음성 아파트키, 자동차 내에서의 다이얼링 등과 같은 실생활의 여러 부분에서 활용 될 수 있는 즉 음성통신 분야에 일대 혁신을 일으킬 것으로 기대된다.

## 참 고 문 헌


[1] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[2] L. R. Rabiner, A. E. Rosenberg, S. E. Levison, "Considerations in Dynamic Time Warping Algorithm for Discrete Word Recognition," *IEEE Trans. ASSP.*, vol. ASSP-26, no. 6, pp. 575-582, Dec. 1978.

[3] Frank Fallside, W. A. Woods, *Computer Speech Processing*. Prentice Hall, 1985.

[4] J. G. Wilpon, C. H. Lee, L. R. Rabiner, "Connected Digit Recognition based on improved Acoustic Resolution," *Computer Speech and Language*, no. 7, pp. 15-26, 1993.

[5] Janet MacIver Baker, "Brief Status Summary for Automatic Speech Recognition at the start of the 80's," *Congress and Exposition Cobo Hall*, Detroit Feb. pp. 25-29, 1980.

[6] 박기영, 신유식, 김종교, "화자 종속 한국어 숫자음 음성 인식 다이얼링 시스템," 대한전자공학회지, 제36권, T편, 제2호, pp. 56-62, 1999,. 6

[7] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[8] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, vol.

50, no. 2, pp. 637-655, 1971.

[9] Kazuo Nakata, *Fundamentals of Speech Signal Processing*. Academic Press, 1989.

[10] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257-285, Feb. 1989.

[11] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization," *IEEE Trans. on Comm.*, vol. COM-28, no. 1, pp. 84-95, Jan. 1980.

## 저 자 소 개

崔 炯 箕(正會員)

1971年 1月 10日生. 1995年 2月 전북대학교 전자공학과 졸업. 1997年 2月 전북대학교 대학원 전자공학과 공학석사 학위 취득. 1999年 2月 전북대학교 대학원 전자공학과 박사과정 수료. <주관심분야 : 음성인식, 음성합성>

朴 基 榮(正會員) 第 36卷 T編 第 2號 參照
全州工業大學, 정보통신과, 교수

金 種 玟(正會員) 第 36卷 T編 第 2號 參照
全北大學校 電氣電子制御工學部, 교수