한국음향학회지 제26권 제6호 pp. 250-258 (2007)

# 중증 장애우용 음성구동 휠체어를 위한 강인한 음성인식 알고리즘

# Robust Speech Recognition Algorithm of Voice Activated Powered Wheelchair for Severely Disabled Person

석 수 영*, 정 현 열**
(Soo-Young Suk*, Hyun-Yeol Chung**)
*산업기술종합연구소 정보기술연구부문 음성처리그룹, 일본 **영남대학교 정보통신공학과
(접수일자: 2007년 6월 7일; 수정일자: 2007년 8월 3일 채택일자: 2007년 8월 10일)

현재의 음성인식 기술은 하드웨어 기술의 발전과 더불어 여러 분야에 응용되고 있지만 음성구동 휠체어와 같은 고신뢰성이 요구되는 응용분야에서는 아직도 그 성능이 불충분하다. 실 환경에서 음성을 통해 안전하게 휠체어를 제어하기 위해서는 도로의 소음 등과 같은 주변잡음의 영향에 의한 음성인식 성능의 저하, 사용자의 기침소리나 숨소리 등과 같은 비음성 입력시의 오동작, 명령어의 불명확한 발성과 일반인과는 다른 발성 속도 및 발성 주파수 등을 고려한 인식시스템이 필요하다. 이를 위하여 본 논문에서는 비음성 입력시의 오동작을 방지하기 위해 인식기의 전처리단에서 YIN 기본주파수 추출방법을 적용한 후 프레임 별 신뢰도에 기반한 고정도로 음성/비음성을 판별할 수 있는 방법을 제안하고, 불명확한 발성에 대한 인식 성능 향상을 위해 화자 적응화 방법 및 개인적인 발성 변이를 표현할 수 있는 다중 후보 단어사전을 구성하여 인식성능 제고를 도모하였다. 잡음이 포함된 실 환경하에서 수집한 데이터를 대상으로 인식실험을 수행한 결과 기존의 켑스트럼 방법에서는 오류 없이 비음성을 찾아내는 재현율은 62%로 나타났으나 본 논문에서 제안한 YIN방법에 기반을 둔 신뢰도 측정방법에서는 95.1%를 나타나 우수한 성능을 나타내었다. 실 환경에서 수집된 2211개의 불명확한 발성을 대상으로 인식실험을 수행한 결과 2000상태 16 혼합수 HMnet 모델을 이용한 경우 인식률이 78.6%로 나타났으나 MAP적응화 방법 및 다중 후보 인식사전을 적용한 결과 99.5%의 인식 성능을 나타내어 제안한 방법의 유효성을 확인할 수 있었다.

**핵심용어:** 음성인식, 비음성 거절, 음성구동 전동 휠체어, 불명확한 발성

**투고분야:** 음성처리 분야 (2.5)

Current speech recognition technology has achieved high performance with the development of hardware devices, however it is insufficient for some applications where high reliability is required, such as voice control of powered wheelchairs for disabled persons. For the system which aims to operate powered wheelchairs safely by voice in real environment, we need to consider that non-voice commands such as user's coughing, breathing, and spark-like mechanical noise should be rejected and the wheelchair system need to recognize the speech commands affected by disability, which contains specific pronunciation speed and frequency. In this paper, we propose non-voice rejection method to perform voice/non-voice classification using both YIN based fundamental frequency (F0) extraction and reliability in preprocessing. We adopted a multi-template dictionary and acoustic modeling based speaker adaptation to cope with the pronunciation variation of inarticulately uttered speech. From the recognition tests conducted with the data collected in real environment, proposed YIN based fundamental extraction showed recall-precision rate of 95.1% better than that of 62% by cepstrum based method. Recognition test by a new system applied with multi-template dictionary and MAP adaptation also showed much higher accuracy of 99.5% than that of 78.6% by baseline system.

*Key words:* Speech recognition, Non-voice rejection, Voice activated powered wheelchair, Inarticulate speech

*ASK subject classification:* Speech Signal Processing (2.5)

책임저자: 정 현 열 (hychung@yu.ac.kr)
712-749 경북 경산시 대동 214-1 영남대학교 전자정보공학부
(전화: 053-814-2496; 팩스: 053-814-5713)

# I. 서 론

장애우의 기본적인 이동수단인 전동휠체어는 그 편이성으로 인해 이용이 나날이 증가되고 있다. 특히, 중증 장애우의 경우 자신의 의지로 움직일 수 있는 유일한 이동수단으로 간주되고 있다. 또한, 최근 전동 휠체어는 기계적 성능 향상과 더불어 주변환경의 위험 분석을 위한 센서 기술, 영상 처리기술이 속속 개발, 적용되어 그 성능이 나날이 향상되고 있다 [1]. 이와 더불어 개인적 장애에 따른 다양한 입력장치에 대한 필요성이 증가되고 있으며, 그 중에서도 음성인식 장치는 조이스틱을 사용하기 힘든 소아마비 장애우나 근육의 움직임이 부자연스러운 중증 장애우를 위한 기본적인 입력장치로 각광받고 있다.

최근 개발된 음성구동 전동휠체어 시스템으로는 기존의 상용 음성인식 하드웨어를 이용한 휠체어의 예를 들 수 있다 [2]. 이 시스템은 명확한 발성이 가능한 장애우를 대상으로 하고 있으며 잡음이 적은 환경에서의 사용을 전제로 하고 있으므로 이를 실 환경에서 적용하기 위해서는 잡음 환경에서의 인식성능의 저하, 음성 입력의 오류, 사용자의 기침소리나 숨소리 등과 같은 비음성 입력으로부터의 인식오류, 휠체어 사용자가 아닌 주변인 음성입력 억제 등을 고려할 필요가 있다.

중증 장애우의 경우에는 발성이 불명확하고 발성의 속도 및 주파수 등이 일반인과 달라 인식성능의 저하를 초래할 수 있으며 위험상황에 대한 대처가 힘들어 보다 신뢰성 있는 시스템을 요구하고 있다. 그러나 현재까지 개발된 일반적인 전동휠체어는 위에서 열거한 이유들 때문에 자신의 의지로 조작하기 어려움이 있어 저속이더라도 제어가 가능한 휠체어에 대한 요구가 강하다.

일반적으로 실용화된 음성인식 시스템의 입력방식은 음성구동 네비게이션 시스템에서와 같이 버튼을 누른 후 수초 이내에 입력되는 명령어를 음성구간검출 장치를 통해 입력하고 있으나, 장애우가 이용하는 휠체어 시스템은 연속적인 명령어 입력이 가능하고 버튼 사용의 불편함을 배제시켜야 한다. 또한, 연속적인 입력이 가능하게 할 경우에는 그림 1에서와 같이 제어를 위한 음성 명령어 이외에도 다양한 신호가 마이크로 입력되며 이에 대해서도 고려해야 한다.

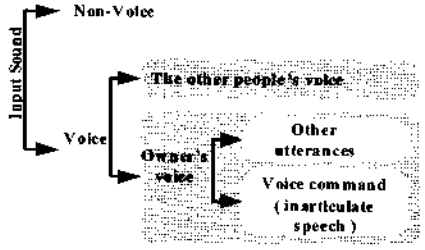



그림 1. 입력 신호의 분류
Fig. 1. Classification of input signals.

음성구간 검출 방법으로서 에너지와 영교차율을 이용하는 경우에는 음성뿐만 아니라 휠체어의 기계음, 사용자의 숨소리, 기침소리 같은 돌발성 잡음도 입력되게 된다. 이와 같은 비음성 입력을 억제하기 위해 에너지와 영교차율의 문턱값을 조정하는 경우 비음성뿐만 아니라 미약하게 발성된 음성도 검출되지 않을 가능성이 높아지게 되며, 이는 반드시 정지해야 하거나 움직여야 할 상황에서 입력음성이 검출 되지 않아 사고로 연결될 수 있다. 따라서 미약하게 발성된 음성의 경우 문턱값에 의한 탈락이 발생하지 않도록 고려할 필요가 있다. 이를 위해서는 입력 문턱 값을 설정한 후 입력된 음성구간에 대해 F0 성분을 추출하여 그 신뢰성 분석을 통해 음성/비음성을 판별하는 방법을 고려할 수 있다.

한편, 중증 장애우의 경우 불명확한 명령어의 발성으로 말미암아 인식성능의 저하가 예상된다. 이 때문에 인식성능의 향상을 위해서는 화자 적응화 방법 및 개인적인 발성 변이를 표현할 수 있는 단어사전의 구성도 필요하다. 또한, 휠체어의 안전한 동작을 위해서는 입력된 음성이 휠체어 사용자의 음성인지 주변인의 음성인지를 판단하여 사용자의 음성에 의해서만 동작하도록 하기 위한 장치, 사용자가 발성한 음성인 경우에도 휠체어 동작을 위한 명령어 인지를 판별할 필요가 있다 [3,4].

본 논문에서는 위에 열거한 연구내용 중 실 환경에서 효과적으로 동작하는 시스템 구현을 위해 신뢰성 높은 음성/비음성 구간 검출 방법과 중증 장애우의 불명확한 발성 및 속도 등을 고려하여 작성된 음성 데이터베이스와 이를 도입하여 개발된 기본적인 음성구동 휠체어 시스템에 대해 간략히 소개하고 새로운 시스템을 이용한 성능평가 결과에 대해 기술한다.

# II. 음성/비음성 분별 장치 및 다중 후보 인식 사전의 구성

## 2.1. YIN을 이용한 음성/비음성 분별장치

실 환경의 음성인식 장치의 경우에는 마이크를 통해 입력된 기침소리, 숨소리, 돌발성 잡음 등과 같은 다양한 비음성 부분을 전처리 단계에서 적절히 억제할 필요가 있

다. 현재까지의 음성/비음성 분별 알고리즘은 음성의 기본주파수, 영교차율, 에너지를 복합적으로 이용하는 방법 [5]과 피치 (F0)의 연속성을 이용한 방법 [6], 뉴럴 네트워크를 이용한 방법 [7] 및 확률적 모델을 이용한 방법 [8]등이 소개되고 있다. 이들 방법 중 대표적인 방법으로는 입력신호의 영역 중 특정 프레임의 F0 정보를 이용하여 음성/비음성 판별에 이용하는 방법이 있다.

그림 2(b)의 예에서와 같이 음성구간검출기로부터 출력된 신호 중 에너지가 가장 큰 하나의 프레임을 검증 대상으로 선택한 후 해당 프레임에서의 F0 정보를 추출하여 주파수가 음성영역 내에 속하는 지를 판단하는 방법이다. 이 방법은 계산량이 적은 장점이 있지만 그림 3의 예에서와 같이 에너지가 최대인 하나의 프레임을 대상으로 검증을 수행함으로 비음성에서도 음성 주파수가 검출되는 경우가 발생하여 검증 성능이 떨어지는 단점이 있다.

또한 입력된 신호의 영역에서 주파수 연결성을 확인하여 음성/비음성 판별을 수행할 수 있으나, 그림 2 (c)의 예에서 보는 바와 같이 음성입력의 경우에도 F0 추출이 완전하지 않아 연결성이 떨어지는 프레임이 나타날 수 있다. 따라서 연결성을 이용하는 방법에서는 식(1)에서와 같이 이전 프레임과의 연결성의 임계값을 결정하기 어렵고, 잡음환경 등의 입력환경이 변화하는 경우에는 임계값을 재설정해야 하는 단점이 있다.

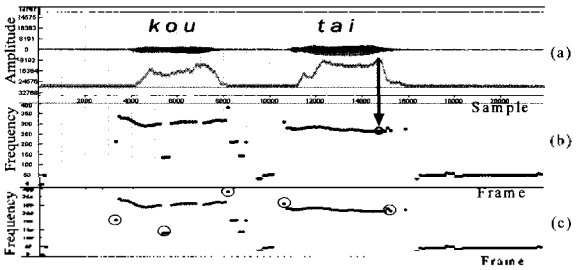



그림 2. 기존의 음성입력에 대한 음성/비음성 분별 척도 (a) 음성파형 (b) 최대에너지에서의 주파수 (c) 추출된 주파수의 연결성

Fig. 2. Usual voice/non-voice classification measure for voice input. (a) Waveform. (b) Frequency with maximum energy. (c) Continuity of extracted frequency.

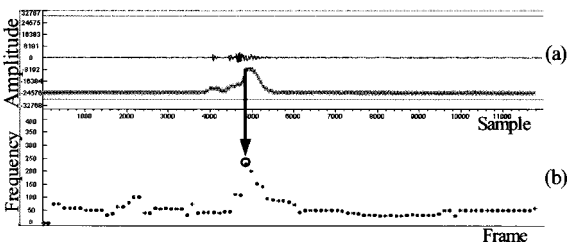



그림 3. 비음성입력에서의 기존의 음성/비음성 분별 척도 (a) 비음성파형 (b) 최대에너지에서의 주파수

Fig. 3. Usualvoice/non-voice classification measure for non-voice input. (a) Waveform. (b) Frequency with maximum energy.

$$|F_{0th}(i)-F_{0th}(i-1)|\le TH \tag{1}$$

음성구동 전동 휠체어 시스템과 같이 한정된 명령어 인식을 위한 고정도 음성/비음성 분별을 위해서는 이와 같은 문제점을 해결한 새로운 방법이 필요하다. 이를 위하여 본 논문에서는 F0 추출과정 중 프레임별 신뢰도 검증을 수행한 후 전체적으로 신뢰성 있는 프레임의 비율을 통해 음성/비음성 구간을 검출하는 새로운 방법을 제안하고자 한다.

F0 검출방법은 시간 영역에서의 검출방법으로는 평균 크기 차분함수 (average magnitude difference function), 평균 제곱 차분 함수 (average squared difference function), 유사 자기상관 함수 (similar autocorrelation methods) 등을 이용하는 방법들이 있으며, 주파수 영역에서는 켑스트럼 분석을 통한 검출방법이 있다. 본 논문에서는 이와 같은 F0 검출 알고리즘 중에서 가장 정확성이 높은 것으로 알려진 YIN 방법을 이용하기로 한다. 이 방법은 de Cheveigne에 의해 제안되었으며 부가적인 파라미터의 수가 적고, 세부적인 조정이 필요 없는 장점이 있다 [9]. 이하 이 방법에 대해 간략한다.

이산 신호 $x_t$ 의 자기상관함수는 식 (2)와 같다.

$$r_t(\tau)=\sum_{j=t+1}^{t+W} x_j x_{j+\tau} \tag{2}$$

여기서 $r_t(\tau)$는 시간 $t$, 지연 성분 $\tau$ 에서의 자기상관 함수이며, $W$ 는 윈도우 사이즈를 나타낸다. YIN방법은 바이어스 값에 영향을 받는 식(2)의 자기상관 함수대신에 차분함수를 이용한다.

$$d_t(\tau)=\sum_{j=t-\tau/2-W/2}^{t-\tau/2+W/2}(x_j-x_{j+\tau})^2 \tag{3}$$

식 (3)의 $d_t(\tau)$ 차분함수가 0이 되는 $\tau$ 값을 구하기 위해 $\tau$ 값이 증가함에 따라 윈도우 폭이 감소한다. 즉 그림 2 (a)에서 보는 바와 같이 높은 $\tau$ 값에서는 $d_t(\tau)$ 크기가 전체적으로 일정 하게 감소하게 되어 F0의 고차 성분이 검출되는 오류를 줄일 수 있는 효과가 있다. 또한 차분함수에서는 신호가 0에서 시작함으로 인해 비정상적인 첫 번째 굴곡점이 나타날 수 있으므로, 검색시 비정상적인 굴곡점을 배제할 목적으로 검색영역을 설정해서 F0 추출

되는 오류를 방지하고 있다. 그러나, 검색영역의 설정에 있어서도 불명확한 주기 때문에 정확한 설정이 어렵다. 이를 해결하기 위해 YIN 방법은 식 (4)와 같이 누적 평균 차분함수를 사용한다.

$$d'_t(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) \Big/ \left[ (1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{otherwise} \end{cases} \quad (4)$$

그림 4 (b)의 예에서와 같이 누적평균 차분함수를 이용하는 경우에는 비정상적인 첫번째 굴곡점이 검출될 수 있는 오류와 고조파 성분이 검출되는 오류를 줄일 수 있으며, 검색영역을 세부적으로 지정할 필요가 없어 F0 검출 영역을 넓힐 수 있는 장점이 있다.

음성/비음성 분별은 F0 추출과정 중 먼저 각 프레임별로 신뢰성을 판단한 후 전체적으로 신뢰성 있는 프레임의 비율을 이용하여 수행한다. YIN 방법을 이용한 프레임별 신뢰성 검증은 식 (4)의 누적평균 차분 함수에 문턱값 $\alpha$ 를 설정하여 $d'_t(\tau)$ 최소값이 문턱값 이하이면 신뢰성이 있는 프레임으로 판단한다.

여기서 신뢰성 있는 F0 프레임의 비율은 식 (5)와 식 (6)에 나타낸 바와 같이 d 값이 분별 임계값 이상이면 음성으로 판단한다.

$$d = \frac{1}{M} \sum_{i=1}^{M} P_{th}(i) \quad (5)$$

$$P_{th}(i) \begin{cases} 1 & \text{if } F_{\min} \le F_{0th}(i) \le F_{\max} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

여기서 M은 입력 프레임의 총수를 나타내며, 프레임별로 검출된 F0주파수를 미리 설정해둔 음성의 범위(60Hz~800Hz)에 속하는지에 대해 재검사를 수행하여 검출 성능을 향상시킨다. 이때 $F_{\min}$=60Hz and $F_{\max}$=800Hz로 설정하였는데 그 이유는 일반인 보다 발성 주파수가 높은 장애우 음성에 대응하기 위해서 이다.

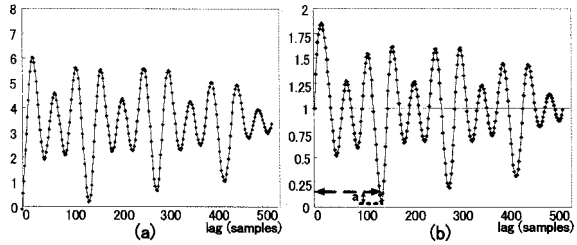



그림 4. (a) 차분함수의 예 (b) 동일한 신호에서의 누적 평균 차분함수의 예
Fig. 4. (a) An example of difference function. (b) Cumulative mean-normalized difference function of the same waveform.

그림 5에 나타낸 바와 같이 음성명령어의 경우 문턱값 이하의 신뢰성 있는 프레임이 음성영역에서 연속적으로 일정값 이상 나타남을 확인할 수 있으며, 그림 6의 비음성의 경우에는 신호의 에너지성분이 큼에도 불구하고 3 프레임에서만 나타남을 확인할 수 있다. 이때 음성 샘플링 주파수는 16kHz이다.

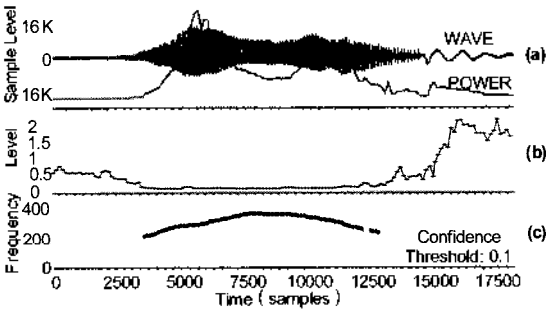



그림 5. (a) 명령어 음성 파형의 예 (b) (a)음성으로 계산된 누적평균 차분함수의 최소값 (c) 신뢰도 검증이 적용된 F0 곡선
Fig. 5. (a) Example of a voice waveform. (b) Minimum value of the cumulative mean-normalized difference function calculated from the waveform in (a). (c) Reliable F0 contour in which the confidence threshold is applied.

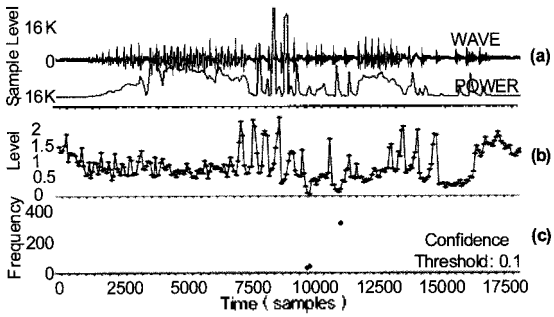



그림 6. (a) 잡음 파형의 예 (b) (a)잡음으로 계산된 누적평균 차분함수의 최소값 (c) 신뢰도 검증이 적용된 F0 곡선
Fig. 6. (a) Example of a noise waveform. (b) Minimum value of cumulative mean-normalized difference function calculated from the waveform in (a). (c) Reliable F0 contour where the confidence threshold is applied.

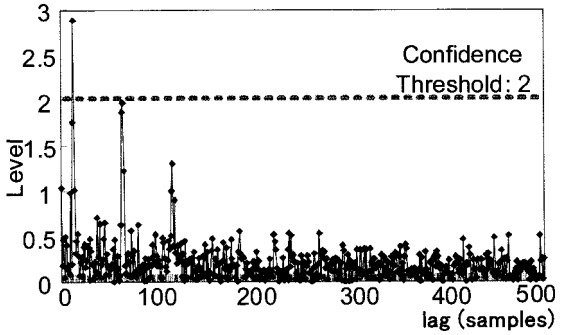



그림 7. 신뢰도 문턱값이 적용된 켑스트럼 신호의 예
Fig. 7. An example of cepstrum signal applied with confidence threshold.

표 1. 불명확한 발성을 위해 작성된 다중 후보 인식사전
Table 1. Multi-template dictionary for inarticulately uttered speech recognition.

| | Command | Template |
|---|---|---|
| Move Left | hidari<br>hidari<br>hidari | h i d a: r i:<br>d a r i q<br>h i h i i d a: r i<br>, … 25 other |
| Move Right | migi | m i g i i, … other 15 |
| Move Forward | mae | m a: a e, … other 13 |
| Move Backward | koutai | k o u t a i, … other 3 |
| Stop | ah | a: , … other 5 |

프레임별 신뢰성 검증은 켑스트럼 방법에서도 적용할 수 있으며, 그 방법은 F0 검출을 위해 켑스트럼 영역의 최대값 부분을 선택할 때 신뢰도 임계값을 적용하여 임계값 이상이면 신뢰성 있는 프레임으로 판단한다.

## 2.2. 다중 후보 인식사전

중증 장애우를 위한 음성명령어의 선택은 장애의 정도와 개개인의 특성에 따라 발음이 쉽고 명료하게 발성할 수 있는 단어로 한정할 필요가 있다. 이를 위해 명령어로 사용 가능한 12개의 단어 가운데 개인별 테스트를 통해 발성이 쉬운 5개의 단어를 선택하였다. 하지만, 이렇게 선택된 단어의 경우에도 장애우의 음성발성의 어려움으로 인해 "hi da ri"의 발성을 "hi hi hi da ri", "i a ri" 등으로 발성하는 경우가 있으며, 이때 일반적인 단어사전의 단어당 발성 후보가 하나로 구성된 경우에는 불명확한 발성에 대응하지 못하는 문제점이 있다. 따라서 본 연구에서는 장애우가 발성한 음성 데이터의 발성 패턴을 분석한 후 다중 후보 인식사전을 구성하기로 하였다. 이를 표 1에 나타낸다.

표1에 나타낸 다중 후보 인식사전을 작성하는 데 있어서는 사전에 녹음한 음성을 대상으로 음소 인식실험을 수행하여 그 결과를 바탕으로 초기 단어 인식사전을 작성한 후, 반복적인 단어인식 실험을 통해 불필요한 후보를 삭제하는 과정을 거쳐 최종 인식사전으로 하였다.

# III. 시스템 개요 및 음성 데이터베이스

## 3.1 음성구동 휠체어 시스템

휠체어 구동을 위한 음성인식 장치는 고정도의 인식성능, 입력장치의 편이성, 빠른 응답, 자유로운 조작 등을 고려한 설계가 필요하다. 그림 8은 II장에서 기술한 내용을 반영한 음성구동 휠체어의 개략도를 나타내고 있으며 이하 시스템 동작에 대해 간략하게 설명한다.

먼저 마이크로부터 입력된 음성신호는 프레임 단위로 분리되어 음성구간 검출, F0추출, Mel Frequency Cepstral Coefficient (MFCC) 특징벡터 추출이 이루어진다. 추출된 F0 성분으로부터 음성/비음성 판별을 수행하게 되며, MFCC특징벡터는 Julian 디코더 [10]를 통해 음성인식이 수행되고, 신뢰성 검증을 통해 인식결과가 명령어인지를 확인하게 된다. 이 때 음향모델은 고정도 인식을 위해 적응화된 Hidden Markov Network (HMnet) 모델을 이용한다 [11]. 최종적으로 음성/비음성 판정결과, 인식 및 신뢰도 결과를 종합하여 동작이 필요한 경우 아날로그 제어기를 통해 전동휠체어를 동작하게 된다.

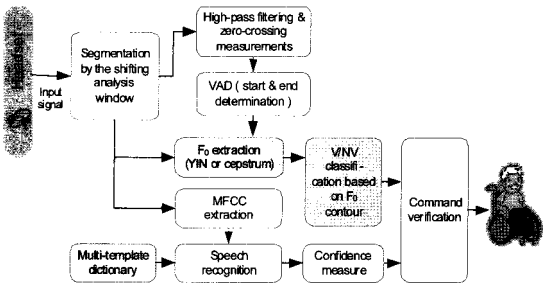



그림 8. 휠체어 구동을 위한 음성인식 시스템의 개략도
Fig. 8. Block diagram of voice activated powered wheelchair system.

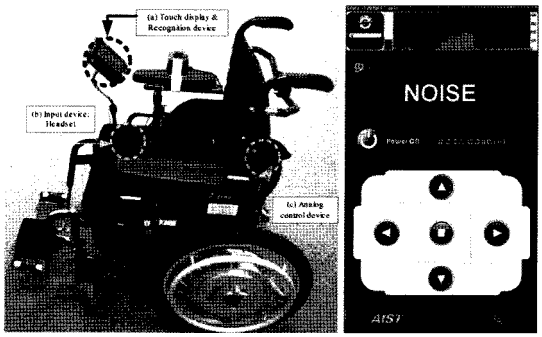



그림 9. 음성구동 휠체어 시제품.
Fig. 9. Developed prototype powered wheelchair.

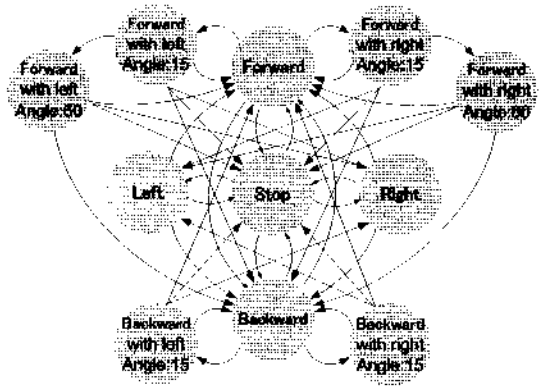



그림 10. 음성명령어에 의한 상태천이도.
Fig. 10. State transition diagram of voice commands.

음성입력을 위한 기본장치로 유선 헤드셋 마이크와 고정용 핀 마이크, 소형 골전도 마이크, 그리고 무선 블루투스 마이크를 이용할 수 있도록 고려되었으며, 그림 9는 실제 개발된 시스템의 한 예를 나타낸다.

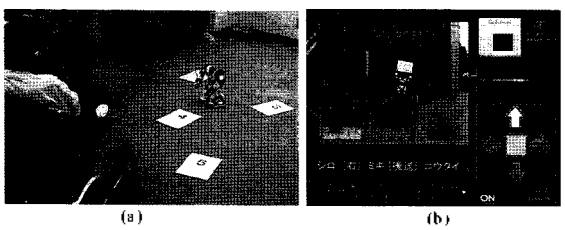
그림 9의 기본 시스템은 마이크 어레이 입력장치 대신 유선 헤드셋 마이크를 사용하였으며, 음성인식 장치 및 사용자 인터페이스를 위해 펜티엄 M 1.2GHz 타블렛 PC에서 동작하도록 하였다. 이때 화면 디자인은 음성입력 상태, 휠체어 동작상태를 사용자가 직관적 확인이 가능하도록 하였고, 전동휠체어의 전체 속도를 변경하기 위해 실내, 저속, 중속, 고속 모드로 변경이 편리하도록 하였다.

음성입력장치를 이용하여 전동휠체어를 조작하는 경우 디지털로 표현된 음성명령어를 아날로그 입력장치인 조이스틱과 같이 각도와 속도를 자유롭게 조작하기에는 어려운 점이 있다. 따라서 보다 더 자유로운 조작을 위해 그림 10과 같이 전후좌우 및 정지의 5개의 상태에 혼합상태 6개를 추가하였다. 예를 들어 "전진" 명령어 이후 "오른쪽" 명령어가 발성되면 15° 각도의 오른쪽 방향으로 전진하게 되고, 다시 "오른쪽" 명령어가 발성되면 60° 각도의 오른쪽 방향으로 전진하게 된다.

### 3.2 음성 데이터베이스

음성인식 시스템의 인식결과는 사용된 데이터베이스에 종속되므로 실용화 시스템을 구현하기 위해서는 실사용 환경에서 수집한 자연스러운 발성 데이터가 필요하다. 특히 중증 장애우를 위한 음성구동 휠체어의 경우에는 발성 환경이 휠체어 조작환경과 유사하면서 자연스러운 발성이 나타날 수 있는 안전한 환경에서 녹음하는 것이 필요하다.

이를 위해 초기 녹음단계에서는 그림 11과 같이 음성명령어로 제어할 수 있는 음성구동 소형로봇 시스템과 그래픽 데모 시스템을 구현하여 음성데이터를 수록하였다. 이때 4개의 마이크를 동시에 착용하였으며, 사용된 마이크



(a) (b)

**그림 11. 음성 녹음 환경 (a) 음성구동 소형로봇 (b) 그래픽 데모 시스템**
Fig. 11. Speech recording environment. (a) Voice operated toy robot system (b) Graphical simulation demo system.

**표 2. 사용된 마이크로폰과 샘플링률**
Table 2. Used microphones and their sampling rates.

| Type | Model | Sampling rate |
|---|---|---|
| Headset | Audio-technica: AT810X | 16 KHz |
| Bone conduction | Sony: ECM-TL1 | 16 KHz |
| Pin | PAVEC: MC-105 | 16 KHz |
| Bluetooth | Sonorix: OBH-0100 | 8 KHz |

**표 3. 데모시스템을 이용한 녹음 데이터의 분석**
Table 3. Analysis of the number of recorded data.

| Type | Voice command | Noise | Owner's other utterances | Other people's utterances |
|---|---|---|---|---|
| Headset | 426 | 65 | 76 | 12 |
| Bone conduction | 405 | 339 | 88 | 286 |
| Pin | 399 | 21 | 90 | 361 |
| Bluetooth | 337 | 22 | 64 | 62 |
| Total | 1567 | 447 | 318 | 721 |

는 헤드셋 (Audio-technica: AT810X), 소형 골전도 (Sony: ECM-TL1), 고정형 핀 (PAVEC: MC-105), 블루투스 헤드셋 (Sonorix: OBH-0100)이다. 블루투스 헤드셋은 전송특성상 8kHz 샘플링률로 녹음하였다.

수집한 음성 데이터의 수는 표 3과 같으며, 헤드셋의 경우 정확히 입력된 음성 샘플수가 426개로써 음성입력 성공률이 가장 높음을 알 수 있었다. 골전도 마이크의 경우에는 돌발성 잡음이 가장 많이 나타났으며, 고정형 마이크의 경우에는 주변 사람들의 음성이 많이 입력됨을 확인할 수 있었다.

다양한 실 환경의 음성 수집을 위한 다음 단계에서는 데모 시스템을 통해 수집된 초기단계의 음성 데이터를 기반으로 음성구동 휠체어 시스템을 구현한 후 집, 도로, 재활센터의 실내와 실외, 공원, 장애우 운동회 등의 환경에서 실제 휠체어 테스트를 겸하여 음성 데이터를 녹음하였다. 수집된 2000개 이상의 음성 데이터는 적응화 및 인식성능 분석을 위해 이용하였다.

## IV. 인식 실험 및 결과

### 4.1 음성/비음성 분별실험

제안된 YIN을 이용한 음성 비음성 분별 성능을 확인하기 위해 표 3의 녹음 데이터를 이용하여 켑스트럼 방법과 비교 실험을 수행하였다. 이때 F0 검출을 위한 샘플링율은 16kHz이며, 윈도우 크기는 25ms, 프레임 이동은 8ms 이다. 실험에 사용된 데이터는 4종류의 마이크로폰으로

표 4. 최고의 재현율-정확도에서 4가지 타입 마이크로폰의 신뢰도 문턱값
Table 4. Confidence threshold analysis of four types of micro-phones with the best recall-precision.

| | Headset | Bone conduction | Pin | Bluetooth |
|---|---|---|---|---|
| Cepstrum | 3 | 2.5 | 1.5 | 2 |
| YIN | 0.05~0.1 | 0.06~0.08 | 0.07~0.1 | 0.08~0.1 |

부터 녹음된 음성명령어 데이터 1567개와 잡음 데이터 447개를 이용하였으며, 그 성능은 정확도와 재현율 곡선으로 나타내었다. 이를 그림 12, 13에 보인다.

그림 12에서와 같이 켑스트럼 방법을 이용한 경우 문턱값을 2.5로 설정할 때 정확도 93%, 재현율 94%를 나타내었으며, YIN방법을 이용한 경우 (그림 13) 문턱값을 0.08로 설정할 때 정확도 99%에 재현율 97%를 나타내어 제안한 방법이 켑스트럼 방법에 비해 성능이 뛰어남을 확인할 수 있다.

일반적인 음성인식 장치와 달리 휠체어 구동을 위한 음성/비음성 분별의 경우, 정지 명령어 발성이 비음성으로 분별되어 반드시 정지해야 하는 경우에 정지하지 못해 사고로 이어질 수 있다. 따라서 음성 명령어의 경우에는 반드시 음성으로 판별할 수 있는 임계값의 설정이 필요하다. 이 경우, 그림 12의 켑스트럼 방법에서는 음성은 반드시 음성으로 판별한 정확도가 100%일 경우 비음성을 찾아낸 재현율은 62%이나, 그림 13의 YIN방법에서는 정확도 100%일 때 재현율 95.1%를 나타내어 성능이 우수함을 알 수 있다.

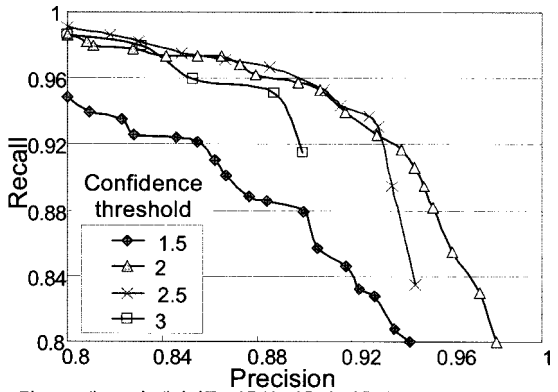



그림 12. 켑스트럼 데이터를 이용한 비음성 검출의 재현율-정확도 곡선
Fig. 12. Recall-precision curve for non-voice classification using cepstrum method.

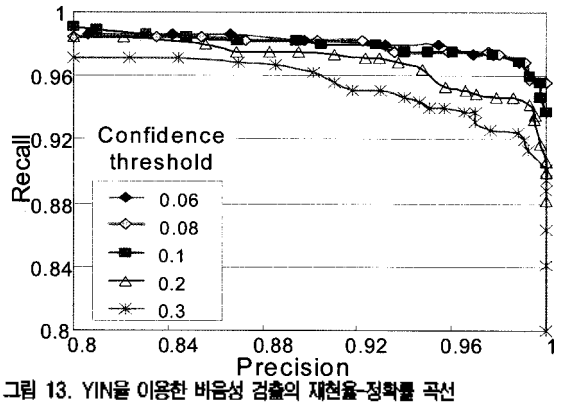



그림 13. YIN을 이용한 비음성 검출의 재현율-정확률 곡선
Fig. 13. Recall-precision curve for non-voice classification using YIN.

마이크로폰 종류에 따른 최적의 신뢰도 문턱값을 결정함에 있어서도 켑스트럼 방법의 경우에는 각 마이크로폰마다 문턱값의 세부 조정이 필요하나, YIN 방법의 경우에는 문턱값 0.08에서 사용 마이크로폰에 관계없이 안정적으로 동작함을 확인할 수 있었다 (표4 참조).

## 4.2. 불명확한 발성에 대한 인식 실험

음성구동 휠체어 시스템의 기본적인 성능과 적응화된 음향모델 및 다중 후보 인식사전의 유효성을 확인하기 위해 인식실험을 수행하였다. 실험을 위해 사용된 데이터와 분석 조건은 표 5와 같다. 표 5의 JNAS 데이터베이스는 일본어 대어휘 연속음성인식을 위해 작성된 남녀 각 153명의 신문기사 낭독체 연속음성 데이터베이스로써 기본 음향모델 생성을 위해 사용되었다.

화자독립 음향모델을 이용한 기본 인식결과와 MAP 방법을 이용하여 적응화된 음향모델의 인식결과를 표 6에 나타내었다. 장애우 운동회 및 다양한 잡음이 포함된 실외 환경 (SNR 10dB이상)에서 녹음된 데이터를 이용한 단어인식 실험에서 1000상태 HMnet 모델을 이용한 경우, 인식사전의 단어 수가 5개임에도 불구하고 단어인식률이 88.2%로 낮게 나타났다. 이 원인은 화자의 특성 및 발성변이를 고려하지 않아 발생한 것으로 판단되었다.

화자변이를 고려한 다중 후보 인식사전의 성능을 확인하기 위해 인식실험을 수행한 결과 표 7에서와 같이 2000상태 HMnet 모델을 이용한 경우, 16혼합의 화자독

표 5. 음성분석 조건 및 데이터
Table 5. Analysis condition and data.

| | |
|---|---|
| Preprocessing | Sampling: 16 KHz, 16 bits<br>Window: 16 ms Hamming<br>Frame shift: 10 ms |
| Features | 24 order of MFCC & 2 order of power |
| Based DB | JNAS Female 10390 sentence |
| Type of microphones | Headset |
| Recorded data | 2211 words (12times)<br>(include data recorded in field and noise environments) |
| Adaptation data | 96 words from 5 times of recorded data |
| Test data | 1334 words from 7 times of recorded data |

표 6. 적응화 단어 수에 따른 MAP 적응화된 모델의 인식성능
Table 6. Recognition accuracies according to the number of adapted words.

| Models | Number of Adapted words | | | | |
|---|---|---|---|---|---|
| | BASE LINE | 12 | 24 | 48 | 96 |
| Monophone | 78.9 | 95.0 | 95.2 | 96.6 | 96.6 |
| Triphone | 51.8 | 76.5 | 91.3 | 97.9 | 99.0 |
| 1000-state HMnet | 88.2 | 97.4 | 98.0 | 99.0 | 99.0 |
| 2000-state HMnet | 78.6 | 97.2 | 98.6 | 99.3 | 99.4 |

표 7. 다중 후보 인식사전을 적용한 적응화된 2000 상태 HMnet 모델의 인식성능
Table 7. Recognition accuracy of 2000 state adapted HMnet model with multi-template dictionary.

| Mix. | Baseline | Multi-Tem. | Adapt. | Multi-Tem. & Adapt |
|---|---|---|---|---|
| 1 | 61.4 | 94.2 | 97.8 | 98.4 |
| 2 | 78.4 | 95.5 | 98.8 | 99.1 |
| 4 | 77.9 | 94.6 | 98.7 | 99.2 |
| 8 | 80.4 | 94.3 | 98.8 | 99.3 |
| 16 | 78.6 | 93.8 | 98.6 | 99.5 |
| 32 | 75.1 | 91.2 | 98.4 | 99.4 |

립 모델을 이용한 경우 인식률이 78.6%로 나타났으나 여기에 다중후보 인식사전을 적용한 결과 93.8%로 향상되어 71%의 오류율이 감소됨을 확인할 수 있었다.

장애우의 경우 적응화 단어발성이 쉽지 않은 점을 고려하여 24 단어를 이용하여 적응화를 수행한 모델의 결과 98.6%의 인식률을 나타내었으며 여기에 다중 후보 인식사전을 추가적으로 적용한 결과 인식률이 99.5%로 나타나 실용화에 충분한 성능을 보임을 알 수 있었다. 이때 오류율 감소는 64%로 나타났다.

## V. 결 론

본 논문에서는 중증 장애우들이 실 환경에서 음성인식 장치를 이용하여 안전하게 휠체어를 제어할 수 있도록 개발한 음성구동 휠체어 시스템의 성능향상을 위하여 인식기의 전처리단에서 고정도 F0추출과 신뢰도 판별을 수행하여 음성/비음성 분별을 수행할 수 있는 방법을 제안하고, 불명확한 발성에 대한 음성인식 성능 향상을 위해 화자 적응화 방법 및 개인적인 발성 변이를 표현할 수 있는 다중 후보 단어 사전을 작성하여 이용하였다.

잡음이 포함된 실 환경하에서 수집한 데이터를 대상으로 인식실험을 수행한 결과 기존의 켑스트럼 방법에서는 오류 없이 비음성을 찾아낸 재현율은 62%로 나타났으나 본 논문에서 제안한YIN방법에 기반을 둔 신뢰도 측정방법에서는 95.1%의 우수한 성능을 나타내었다. 실 환경에서 수록된 2211개의 불명확한 발성을 대상으로 인식실험을 수행한 결과 2000상태 16 혼합수 HMnet 모델을 이용하고 MAP적응화 방법 및 다중 후보 인식사전을 적용한 결과 99.5%의 인식 성능을 나타내어 제안한 방법의 유효성을 확인할 수 있었다.

## 감사의 글

이 논문은 2006학년도 영남대학교 학술연구조성비 지원에 의한 것임

## 참 고 문 헌

1. D. Ding, R.A. Cooper, "Electric Powered Wheelchairs," in IEEE Trans. Control Systems Magazine, 25 22-34, 2005.
2. 송병설, 이정현, 박정제, 박희준, 김명남 "화자 독립 방식의 음성인식 칩 및 무선마이크를 이용한 전동 휠체어의 구현" 센서공학회 논문집, 13 (1) 20-26, 2004.
3. A. Sasou, H. Kojima, "Multi-Channel Speech Input System for a Wheelchair," in Proc. Acoust. Soc. Japan, 2006.
4. K. Sadohara, S.W. Lee and H. Kojima, "Topic Segmentation Using Kernel Principal Component Analysis for Sub-Phonetic Segments," Technical Report of IEICE, AI 2004-77, 37-41, 2005.
5. J. Rouat, Y. C. Liu and D. Morrisette, "A Pitch Determination and Voiced/Unvoiced Decision Algorithm for Noisy Speech," in Speech Communication, 21, 1997.
6. S. Ahmadi, S. S. Andreas, "Cepstrum-based Pitch Detection using a New Statistical V/UV Classification Algorithm," in IEEE Trans. Speech Audio Processing, 7 (3) 333-339, 1999.
7. H. Miyabayashi, T. Funada, "Pitch extraction and voiced/unvoiced detection of speech by cross-coupling multi-layered neural network with feedback architecture," in Journal of Electronics and Communication of Japan, 80 (9) 48-58, 1998.
8. K. Giridharan, B.Y. Smolenski and R.E. Yantorno, "Statistical And Model Based Approach To Unvoiced Speech Detection," in Proc. ISPACS, 816-821, 2004.
9. A. de Cheveigne, H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," in Journal of the Acoustic Society of the America, 111, 2002.
10. A. Lee, T. Kawahara and K. Shikano, "Julius - an Open Source Real-time Large Vocabulary Recognition Engine," in Proc. European Conference on Speech Communication and Technology, 1691 1694, 2001.
11. S.Y. SUK, S.W. Lee, H. Kojima and S. Makino, "Multi-mixture based PDT-SSS Algorithm for Extension of HMNet Structure," in Proc. Acoust. Soc. Japan, 2005

## 저자 약력


• **석 수 영 (Soo-Young Suk)**

1998년 2월: 계명대학교 물리학과 (이학사)
2000년 2월: 영남대학교 멀티미디어 통신공학과 (공학석사)
2004년 2월: 영남대학교 정보통신 공학과 (공학박사)
2004년 3월~2005년 3월: (일본) 동북대학교 전자통신공학과 연구원 박사후 과정
2005.4~현재: (일본) AIST 음성정보처리 연구그룹 연구원
※ 관심분야: 음성인식, 음성처리, 멀티미디어 검색

• **정 현 열 (Hyun-Yeol Chung)**

1989년: 일본 동북대학교 정보공학과(공학박사)
1989년3월~현재: 영남대학교 전자정보공학부 교수
1992년7월~1993년 7월: 미국 CMU Robotics 연구소 객원연구원
2000년6월~2000년 8월: 미국 Qualcomm Inc. 수석 엔지니어
※ 관심분야: 음성인식, 화자인식, 음성합성 및 DSP 응용분야