工學碩士 學位論文

DTW를 이용한 선박 음성조타명령 인식시스템의 구현

Implementation of Vessel's Steering Command Recognition System using DTW

指導敎授 辛 沃 根

2006年 2月

韓國海洋大學校 大學院

컴퓨터工學科

南 彦 圭

本 論文을 南彦圭의 工學碩士 學位論文으로 認准함

委員長　工學博士　　柳 吉 洙　印

委　員　工學博士　　金 載 薰　印

委　員　工學博士　　辛 沃 根　印

2005年　12月

韓國海洋大學校 大學院

차 례

제 1 장 서 론 ····· 1

제 2 장 음성인식을 위한 전처리 ····· 2

2.1 음성인식시스템의 구성 ····· 2

2.2 음성의 생성 및 지각 ····· 3

2.3 음성 구간 추출 ····· 5

2.4 특징벡터추출 ····· 9

제 3 장 DTW를 이용한 단어 인식 ····· 17

3.1 DTW ····· 18

3.2 DTW를 이용한 단어 인식 ····· 25

제 4 장 문법에 의한 조타명령 인식 ····· 29

4.1 조타명령 ····· 29

4.2 조타명령 문법 ····· 32

4.3 조타명령인식을 위한 문법적용 ····· 33

제 5 장 실험 평가 및 결과 ····· 35

5.1 인식 실험 환경 및 음성 데이터 ····· 35

5.2 화자독립 단어 인식 실험 ····· 36

5.3 화자독립 조타명령 인식 실험 ····· 37

5.4 화자종속 조타명령 인식 실험 ····· 40

제 6 장 결론 및 향후 과제 ····· 42

참고 문헌 ····· 43

Implementation of vessel's steering command recognition system based on DTW

Eun Kyu Nam

Department of Computer Engineering, Graduate School,
Korea Maritime University, Busan, Korea

Advised by Prof. Ok Keun Shin

Abstract

In this dissertation, we present an implementation of automatic speech recognition system designed to recognize vessel's steering command. The system first detects the point where the speech starts and ends, and then processes the acoustic signal to produce 13th order MFCC. The recognition of steering command is conducted in 2 steps: the calculation of similarity in word level, and then the selection of the final steering command from the set of feasible commands. The word level similarity between test and reference MFCC vectors is calculated by DTW (Dynamic Time Warping) method. Then a word lattice is made to aid the selection of the steering command. Steering commands considered in this study consist of one to four words chosen from 23 words including numbers. Three kinds of recognition experiments are conducted to test the performance of the

presented system. In the first two cases which are speaker-independent command-level and word-level experiments, the recognition rate was around 70%. In the last experiment, which is speaker-dependent command-level recognition, the recognition rate is more than 97%, which encourages the application of the system in real environment.

제 1 장 서 론

인간의 음성인식 과정을 자동화하고자 하는 자동음성인식(ASR:automatic speech recognition)에 대한 연구가 오래전부터 활발하게 진행되어 왔으며, 근래에는 산업 판도를 바꿀 미래의 10대 기술로 선정되기도 했다[1]. 정보통신 산업이 발달함에 따라 선박에서도 자동 항법시스템, 선박 제어 및 감시 시스템 등 여러 부분에서 다양한 자동화 설비가 이용되고 있으며, 음성 인식 기술을 활용하고자 하는 시도도 있었다[2].

본 논문에서는 음성인식 기술을 해양 관련 시스템에 응용하여, 선박을 운항하는 항해사가 내리는 조타명령을 자동으로 인식하는 시스템을 제안한다. 제안한 음성조타명령 인식기는 조타명령의 특성에 맞추어 소용량 고립단어 인식에 이용되는 DTW를 사용하였으며, 특징 추출의 전단계로 인식률에 많은 영향을 미치는 음성끝점 검출의 정확도를 높이기 위해서 영교차율과 가중 엔트로피법을 적용하였다. 인식기의 후반부에는 조타명령에 대한 문법을 이용한 word lattice와 명령 구성 단어별 인식후보를 적용하여 인식률 향상을 도모하였다.

본 연구에 사용된 음성데이터는 23개의 단어로 구성되며, 각각의 조타명령은 이들 단어를 1~4개 조합하여 만들어진다. 33명의 화자로부터 단어별 발화를 녹음하고, 이를 이용하여 본 연구에서 구현한 시스템으로 조타명령 인식 실험을 수행한 결과, 화자독립의 경우 70%정도의 인식률을 보였으나 화자종속의 경우 97%의 인식률을 보여 실제 환경에 적용될 수 있을 정도의 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 입력된 음성신호의 전처리 과정을, 3장에서는 DTW를 이용한 단어 인식 방법을, 4장에서는 단어 인식 결과를 이용하여 문법적으로 적합한 조타명령인지를 구별해 내는 문법검증 방법을, 5장에서는 인식 실험 평가와 결과에 대해 고찰하고, 마지막으로 6장에서 결론을 맺는다.

제 2 장 음성인식을 위한 전처리

효율적인 음성인식을 위해 일반적으로 선택하는 방법은 먼저 시간영역의 입력 음성신호를 주파수 영역의 신호로 변환한 다음 특징벡터를 추출하여 기준이 되는 특징벡터와 비교하는 것이다. 그러나 인식기에 입력되는 신호에는 인식의 대상이 되는 음성신호만 있는 것이 아니라 묵음구간을 포함하고 있으므로 인식 대상이 되는 음성 구간만을 찾아서 특징벡터를 추출하는 것이 필요하다.

본 장에서는 본 논문에서 구현하는 음성인식 시스템의 전처리에 대해 설명한다. 인간의 음성 생성과 지각과정에 대해 간략하게 서술하고 전체시스템의 구성을 보인 다음 음성인식의 전처리과정으로 생각할 수 있는 음성구간 추출과정과 추출된 음성에서 특징벡터를 계산하는 과정을 차례로 서술한다.

2.1 음성의 생성 및 지각

음성은 허파로부터 성대와 성도를 통과하는 공기의 흐름에 의해 생성되는데, 발성의 관점에서 허파는 공기 공급원으로, 성대는 기본 음파 발생기로, 그리고 성도는 변조기와 공명기로 동작한다[5].

음성은 크게 유성음과 무성음의 두 가지로 구분할 수 있다. 유성음은 성대가 닫혀 있다가 폐가 공기를 밀어내기 시작하면 압력이 커지면서 성대가 진동함으로써 이루어지는 발성이다. 유성음의 진동 주파수는 성대의 길이와 긴장감의 정도에 의해 결정되는데 이 주파수를 피치 주파수라하며 일반적으로 남자의 경우 50Hz에서 400Hz 정도이다. 무성음은 성대가 보통 열려 있는 상태에서 공기가 성도를 자유롭게 통과한 다음 구강 또는 비강 내에서 마찰, 혹은 파열이 일어남으로써 발성된다[6, 7, 8].

음성의 지각에는 귀와 뇌가 관여하게 되는데, 이들 중 귀가 음향신호를 검출하는 방법은 비교적 많이 알려져 있다[5]. 인간의 귀는 외이, 중이, 내이의 세

부분으로 구성된다. 외이는 외부 소리가 중이로 들어가는 통로이며, 중이는 세 개의 뼈 구조로 음성파의 진동을 내이로 전달한다. 내이는 반원형의 캐널(canal), 와우각(cochlea), 그리고 음향 신경으로 구성된다. 와우각은 유체가 차 있고, 모양은 나선형이다. 이 와우각 안에는 머리카락 형태의 유모세포(hair cell)가 있어 전달된 신호에 대한 주파수 분석기로 작동한다[7].

사람의 청각은 음의 크기를 대수적(logarithmic)으로 인지하고 음의 높이 또한 낮은 주파수는 세밀하게, 높은 주파수는 개략적으로 인지하는 특성을 갖는다. 따라서 청각 기능에서 음의 인지 정도는 캡스트럼(cepstrum)으로 표현할 수 있으며 음의 높이를 느끼는 주파수 분해능은 멜(mel) 척도가 적당하다. 멜 척도란 음의 높이를 느끼는 심리적인 정도로서, Stevens 등[5, 7, 23]에 의해서 연구되었으며, 이들은 1000Hz, 60dB의 기준음을 1000Mel로 정의하여 기준음의 1/2높이로 느끼는 음의 주파수는 500Mel, 2배로 느끼는 음의 주파수는 2000Mel로 하였다[1, 3]. 실제 멜 척도로 계산된 음성 스펙트럼은 선형 척도에서 구한 것 보다 음운성을 잘 나타내기 때문에 보다 유용하게 이용되며 특히 음성을 캡스트럼으로 분석할 때 유용하다[7].

2.2 음성인식 시스템의 구성

그림 2.1은 본 논문에서 구성한 음성 인식 시스템의 구조이다. 마이크를 통해 입력되는 음향신호로부터 음성구간을 추출하는 부분은 “음성구간 추출” 으로, 입력된 음성신호를 pre-emphasis, frame blocking, windowing 등의 처리를 시간영역에서 수행한 다음, 주파수영역으로 변환하기 위해 푸리에변환(fourier transform)을 수행한다[3]. 주파수 영역의 신호에 인간의 청각특성을 적용한 멜 필터 뱅크(mel-filter bank)를 적용하여 MFCC 특징벡터를 계산하는 “특징벡터 추출” 이다. 이렇게 얻은 입력신호의 특징벡터를 기준 패턴 특징벡터와 DTW를 이용해서 유사도를 계산하고, 유사성이 높은 후보들을 찾는데 이 과정은 조타명

령(steering command)의 단어단위로 수행되는 “유사도 측정” 과정이다[4]. 마지막으로 이상의 과정에서 얻은 단어 단위의 인식후보들에 조타명령 문법을 적용시킨 “최종인식 결정” 으로 조타명령을 인식하게 된다.

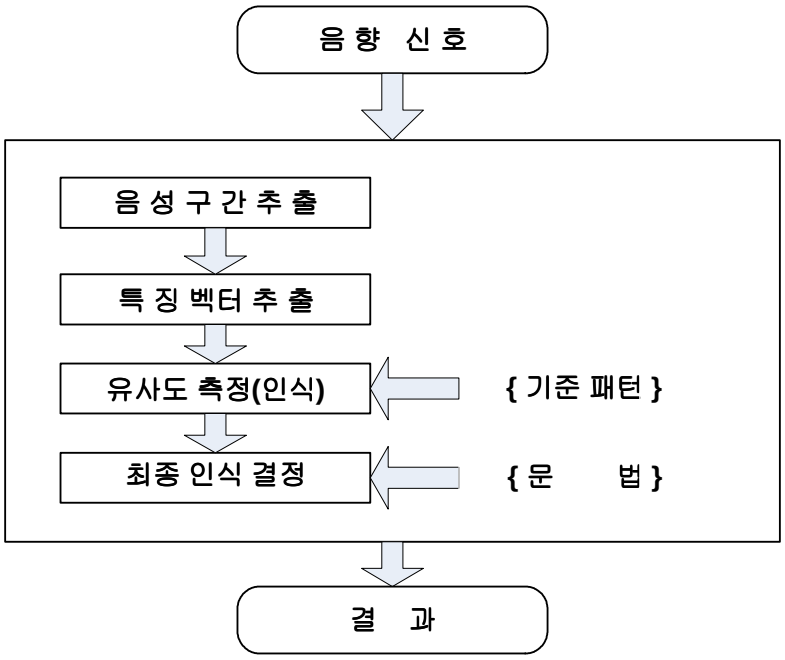


그림 2.1 음성인식 시스템의 구조.

Fig. 2.1 The structure of the speech recognition system.

2.3 음성 구간 추출

인식기에 입력되는 신호에는 인식의 대상이 되는 음성신호만 있는 것이 아니라 묵음구간을 포함하고 있으므로 제일먼저 인식대상이 되는 음성이 있는 구간을 찾아내는 것이 필요하다. 음성 구간을 찾는 대표적 알고리즘인 끝점 검출(end-point detection) 알고리즘에는 시간영역 에너지(short-time energy), 영교차율(zero crossing rate), 자기상관계수(autocorrelation), 피치(pitch), 예측 에러 에너지(prediction error energy), 엔트로피(entropy)를 이용한 방법 등이 있다[5, 8].

본 절에서는 가장 기본적인 끝점 검출 알고리즘인 시간영역 에너지 방법에 대해 설명한 다음, 본 논문에서 이용하는 영교차율과 엔트로피를 이용하는 방법에 대해 기술한다.

(1) 시간영역 에너지

시간영역 에너지를 이용한 방법에서는 일정구간의 시간영역 에너지를 파라미터로 사용한다. 파라미터로 사용하는 구간 에너지 M_n은 식 (2.1)과 같다.

$$M_n = \sum_{m=n-N+1}^{n} | x(m) |^2 w(n-m) \qquad (2.1)$$

여기서 $w(n)$은 프레임 구간 내에서는 1, 그 이외에서는 0의 값을 가지는 사각형의 윈도우다. 해당 구간에서 결정된 에너지 값이 일정 임계값(threshold)을 넘어서면, 음성신호가 있는 구간으로 인식한다. 임계값은 마이크로폰, 주변 환경 등 여러 요인에 의해서 변할 수 있다. 따라서 최적화된 임계값을 구하기 위해서는 다양한 환경에서의 실험이 필요하다.

(2) 영 교차율

영 교차율이란 디지털화된 시간영역의 음성신호에서 두 개의 연속된 샘플의 부호가 바뀌는 비율을 말한다. 해당 구간에서 음성신호의 주파수가 높으면 영 교차율이 증가하고, 주파수가 낮으면 영 교차율이 감소하게 된다. 영 교차율 방법은 유성음 구간의 검출은 물론 시간영역 에너지 방법의 단점인 무성음 구간의 검출에도 효율적인 방법으로 알려져 있다[8].

식 (2.3)에 영 교차율 Z_n을 보였다. 여기서 $\mathrm{sgn}[x(m)]$ 은 식 (2.3)과 같이 $x(m)$값의 부호를 나타내며, $w(n)$은 사각형의 윈도우 함수로 식 (2.4)와 같다.

$$Z_n = \sum_{m=n-N+1}^{n} | \mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\, w(n-m) \quad | \tag{2.2}$$

$$\mathrm{sgn}[x(m)] = \begin{cases} 1 & (x(m) \geq 0) \\ -1 & (x(m) < 0) \end{cases} \tag{2.3}$$

$$w(n) = \begin{cases} \dfrac{1}{2N} & (0 \leq n \leq N-1) \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

(3) 엔트로피를 이용하는 방법

정보이론에서 엔트로피 $H(x)$는 신호 x가 가지는 정보량으로 생각할 수 있으며, 식 (2.5)와 같이 정의된다[9].

$$\begin{aligned} H(x) &= \sum_{i=1}^{N} p(i) \log_2 \left(\frac{1}{p(i)} \right) \\ &= -\sum_{i=1}^{N} p(i) \log_2 p(i) \end{aligned} \tag{2.5}$$

여기서 N은 신호 x가 가지는 상태의 수이며, $p(i)$는 상태 i가 발생할 확률밀도 함수(probability density function)이다.

엔트로피를 이용하여 음성 구간을 추출하는 방법은 기존의 여러 음성구간 추출방법에 비해 잡음이 많은 환경에서 효과적이다[10]. 엔트로피를 이용한 음성구간 추출 방법에서 엔트로피는 신호의 주파수 영역에서 정의되는데, 식 (2.5)의 $p(i)$는 식 (2.6)과 같이 주파수 대역별로 정규화된 에너지(normalized energy)로 정의할 수 있다.

$$p(i) = s(f_i) \,/\, \sum_{k=1}^{N} s(f_k), \qquad i = 1, 2, ..., N \tag{2.6}$$

여기서 $s(f_i)$ 는 주파수 성분 f_i의 스펙트럼 에너지(spectrum energy)이며, N은 FFT(fast fourier transform) bin의 개수이다.

본 논문에서는 엔트로피를 이용한 음성구간 추출의 효율을 높이기 위해 Jia-lin Shen 등[11]이 제안한 아래의 개선된 방법을 이용 했으며 그 과정은 다음과 같다.

첫 번째 단계는 필터링과정으로 식 (2.7)과 같이 250Hz 에서 6000Hz 사이의 주파수 성분들만 취한다. 이 주파수 범위는 대부분의 음성신호들이 있는 구역이다.

$$s(f_i) = 0, \qquad \text{if} \quad f_i < 250\text{Hz} \text{ or } f_i > 6000\text{Hz} \tag{2.7}$$

둘째로, 식 (2.8)과 같이 확률밀도에 최소, 최대 임계값을 벗어나는 성분들의 확률밀도는 0으로 처리한다. 식 (2.8)에서 임계값 δ_1은 통상적인 음성의 주파수 대역을 넘어서는 고주파 대역을 제외하기 위한 것이며, 본 연구에서는 실험을

통해서 0.8로 설정하였다. 또 다른 임계값 δ_2는 대부분의 주파수 대역에 걸쳐 균일한 값을 가지고 있는 화이트 노이즈(white noise) 성분을 제거하기 위한 것이다. 본 연구에서는 입력받은 음향신호의 처음 일정구간(30ms)을 노이즈라 가정하고 이 구간에서 δ_2 값을 구하였다.

$$P(i) = 0, \quad \text{if} \quad P(i) < \delta_2 \quad \text{or} \quad P(i) > \delta_1 \tag{2.8}$$

$$\delta_1 = 0.8 \qquad \delta_2 = \frac{\max\{S(f_i)\} - \min\{S(f_i)\}}{2} + \frac{\sum_{k=1}^{N} S(f_k)}{N}$$

$$k = 1, 2, \ldots, N$$

마지막으로 식 (2.5)의 엔트로피에 주파수 대역별로 가중치를 부여하여 식 (2.9)와 같이 변형시킴으로써 스펙트럼 에너지의 평균적인 크기를 고려할 수 있게 하였다. 주파수 대역별 가중치 $w(i)$는 많은 양의 음성신호로부터 통계적으로 추정한 것으로 식 (2.10)과 같이 스펙트럼 에너지 $s(f_i)$를 각각의 FFT bin 별로 더한 다음 정규화 시킨다. 그림 2.2는 수집한 모든 음성데이터로부터 구한 주파수 대역별 가중치 $w(i)$를 나타낸 그래프이다.

$$H(x) = -\sum_{i=1}^{N} w(i) P(i) \log_2 P(i) \tag{2.9}$$

$$w(i) = \frac{\sum_{l=1}^{L} S(f_i)}{\sum_{l}^{L} \sum_{k=1}^{N} S(f_k)} \qquad i = 1, 2, \ldots, N \tag{2.10}$$

여기서, k는 FFT bin의 색인(index)이며, l은 프레임의 색인, L는 전체 프레

임의 수이다.

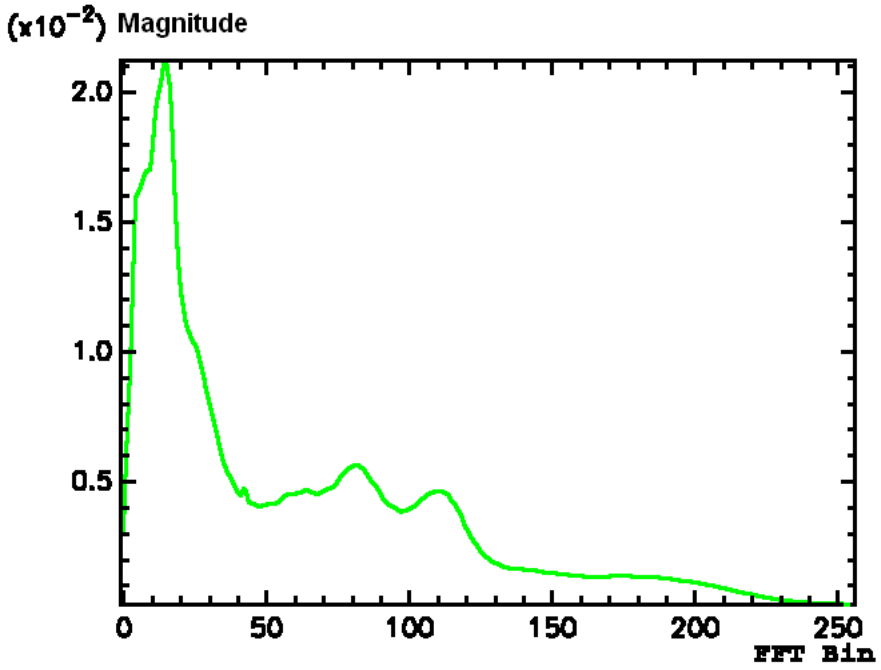


그림 2.2 모든 음성 데이터로부터 구한 가중치 $w(i)$.

Fig. 2.2 Weight factor extracted from all speech signal.

2.4 특징벡터추출

특징벡터추출은 시간 영역의 음성 신호로부터 음향학적 특징을 포함하는 벡터를 추출 하는 과정이며, 이에 관한 연구는 지난 수십 년 간에 걸쳐 진행되어 왔다. 특징벡터를 추출하는 대표적인 방법으로는 음성의 발성과정을 모델링하는 방법과 인간의 음성 인지과정을 모델링하는 방법이 있다[5]. 발성과정을 모델링하는 방법으로는 인간의 성도의 특성을 모델링한 선형 예측코딩(linear prediction coding)이 있으며, 음성 인지과정을 모델링 하는 방법으로는 인간의 청각 특성을 이용한 캡스트럼 계수(cepstrum coefficient), 음성의 포만트

(formant) 특성을 잘 나타내는 LSF(line spectrum frequency), 청각 특성을 이용한 MFCC(mel frequency cepstral coefficient) 등이 있다[12].

본 논문에서는 기존의 연구들에서 많이 활용되어 좋은 결과를 보이고 있는 MFCC와 Delta-MFCC를 특징벡터로 사용하였으며, 이를 추출하는 과정은 그림 2.3과 같다.

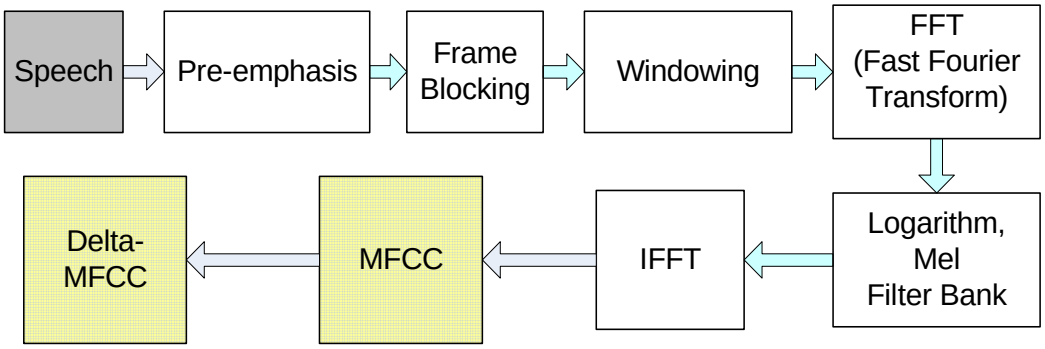


그림 2.3 MFCC 및 Delta-MFCC추출과정.

Fig. 2.3 Block diagram of MFCC and Delta-MFCC extraction.

(1) Pre-emphasis

마이크에 입력된 음성신호는 A/D변환을 거쳐서 디지털 신호 $x(n)$으로 변환된다. 디지털 음성신호는 고대역 통과 특성을 갖는 디지털 pre-emphasis 필터를 거친다. 이 필터는 낮은 주파수의 파워는 감쇄시키고 높은 주파수의 파워는 상대적으로 증폭시키는 효과가 있는데, 이것은 인간의 청각이 1㎑이상의 고주파 영역에 더 민감하다는 사실을 모델링함과 동시에 마이크의 성능에 기인하는 왜곡을 보정하기 위한 것이다. 가장 널리 사용되는 pre-emphasis는 식 (2.11)로 본 연구에서는 a를 0.95로 사용하였다.

$$H(z) = 1 - az^{-1}, \qquad 0.9 \le a \le 1.0 \tag{2.11}$$

Pre-emphasis 네트워크의 출력 $\tilde{s}(n)$은 식 (2.11)을 이용하여 구할 수 있으며, 식 (2.12)와 같다. 여기서 $s(n)$은 입력된 음성신호이다.

$$\tilde{s}(n) = s(n) - as(n-1) \tag{2.12}$$

(2) Frame Blocking

Frame Blocking은 입력음성 신호를 일정한 길이를 갖는 프레임단위의 신호들로 나누는 것으로 blocking된 신호 $X_l(N)$은 식 (2.13)과 같다.

한 프레임의 길이가 N샘플일 때 M이 N보다 클 경우 중첩되지 않게 되며, 일반적으로 M의 값은 $\frac{1}{3}N$정도가 적당하다[5].

$$X_l(N) = \tilde{s}(Ml+n) \quad n=1,2,\ldots,N-1, \quad l=1,2,\ldots,L \tag{2.13}$$

여기서 $\tilde{s}(n)$은 pre-emphasis된 음성신호를, N은 프레임 길이를, M은 인접프레임 사이의 간격을 말한다.

(3) Windowing

다음 단계는 각 프레임의 시작과 끝에서 생기는 신호의 불연속을 최소화하기 위해 각 프레임들에 윈도우(window)를 씌운다. $0 \le n \le N-1$의 범위에서 윈도우를 $w(n)$으로 정의하면, 윈도우 처리된 결과는 식 (2.14)의 신호가 된다.

$$\widetilde{x_l}(n) = x_l(n)\, w(n), \quad 0 \le n \le N-1 \tag{2.14}$$

본 연구에서는 해밍윈도우(hamming window)를 사용했으며 다음의 식 (2.15)와 같다.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \tag{2.15}$$

(4) 푸리에 변환(Fourier Transform)

일반적으로 음성신호는 단기간 정적(short-time stationary)인 성질을 갖는다. 즉, 음성신호의 특성은 짧은 시간 동안에는 크게 변하지 않고 서서히 변한다는 것이다[13]. 단구간 푸리에 변환은 이와 같은 가정 하에서 짧은 구간의 음성신호에 대한 푸리에 변환을 정의한 것이며, 윈도우 함수를 적용한 음성 구간에 대해 주파수 성분을 계산한다. 대부분의 음성인식 시스템은 음성신호가 포함하고 있는 유용한 정보는 유지하고 필요 없는 부분은 제거하기 위하여 특징벡터를 추출하여 인식에 사용하는데, 일반적으로 특징벡터 추출과정은 시간영역보다 음성신호의 유용한 정보를 잘 반영하는 주파수 영역에서 실행된다. 따라서 특징벡터를 추출하기 전에 먼저 시간축상의 음성신호를 주파수 영역으로 변환해야 하며, 이때 많이 사용되는 푸리에 변환은 가장 효율적이고 일반적인 주파수영역 변환 방법이다[14, 15]. 계산의 효율을 위해 식 (2.16)에서 보인 FFT를 이용하여 주파수 영역으로 변환을 행한다.

$$X(k) = \sum_{n=0}^{N-1} \widetilde{x}(n)\, W_N^{kn}, \qquad W_N = e^{-j\left(\frac{2\pi}{N}\right)} \tag{2.16}$$

(5) 멜 스케일 필터(Mel Scale Filter)

Zwicker에[16] 의하면, 인간이 느끼는 주파수는 물리적인 주파수의 로그 함수 형태(bark scale)로 변별력을 갖는다고 한다. 그림 2.4는 물리적인 주파수와 청각인식 주파수와의 관계를 나타내고 있다[23]. 이 그림에서 보는 것처럼 1000Hz 이하의 물리 주파수와 청각 인식 주파수는 비교적 선형적이며, 그 이상의 주파수에서는 비선형적인 특징을 나타낸다[14, 16, 23]. 그림 2.4에 나타낸 변환함수에 의해 변환된 주파수를 멜 주파수(mel frequency)라고 한다.

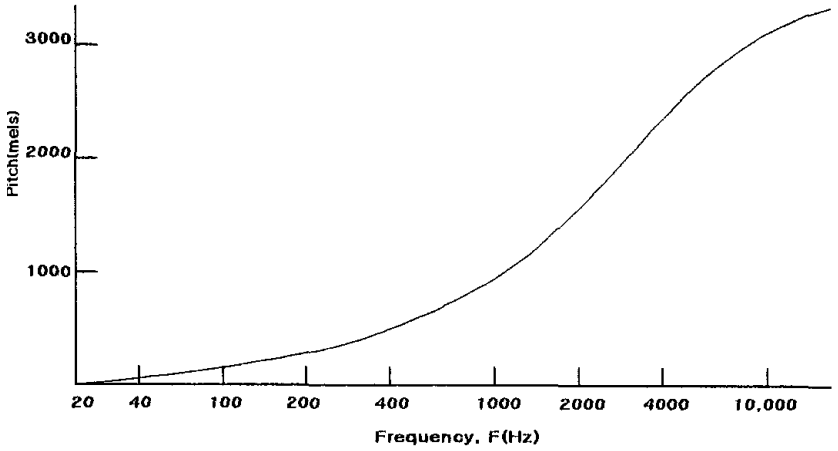


그림 2.4 실제 음성 주파수와 청각 인식 주파수의 관계.
Fig. 2.4 The relation between real speech frequency and auditory frequency.

French 등[24] 은 멜 스케일의 주파수를 이용하여 특징벡터를 추출하는 방법에서 그림 2.5와 같은 삼각필터로 구성된 멜 스케일 필터를 제안하였다. 그림에서 볼 수 있는 것과 같이 1000Hz 이하에서는 선형적인 특성을 보존하기 위해 일정간격으로 균일하게 정렬된 필터들로 구성되고, 그 이상의 주파수 대역에서

는 비선형적인 특성을 반영하기 위해 대역폭이 넓어지는 모양의 필터들로 구성된다.

식 (2.16)의 푸리에 변환으로 얻어진 주파수 영역의 신호에서 파워를 구한 다음, 아래의 그림에 보인 멜 필터뱅크(filter bank)를 적용하여 멜 스케일 파워 스펙트럼 $Y(m)$을 구한다. 이 과정을 식 (2.17)에 보이며, 여기서 m은 필터뱅크의 색인이다.

$$Y(m) = \sum_{k=0}^{\frac{N}{2}} |X(k)|^2 H_m(k) \tag{2.17}$$

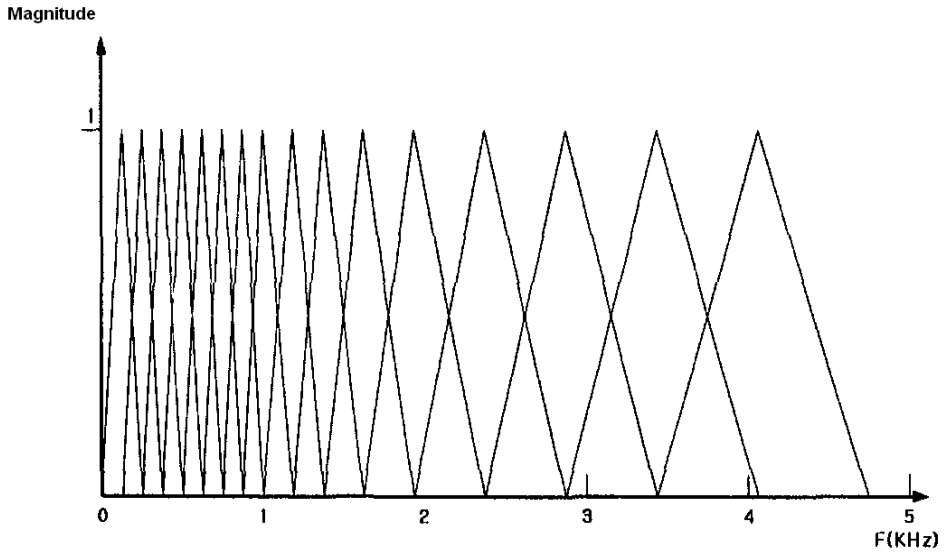


그림 2.5 멜 필터 뱅크 주파수 응답과 멜 주파수 간격에 의한 격자.

Fig. 2.5 A filter-bank frequency response with bandwidth and spacing determined by a constant mel frequency interval.

(6) IFFT(Inverse Fast Fourier Transform)

MFCC를 구하기 위한 마지막 과정은 주파수 상태에 있던 값들을 시간 영역으로 변환해야 한다. 필터뱅크에서 얻어진 출력 $Y(m)$에 로가리즘(logarithm)을 취하고 Inverse DFT를 수행하여 주파수 영역의 값들을 시간 영역의 값들로 변환하여 특징벡터를 구한다.

이 과정에서 IDFT연산은 대칭적인 실수인 로그파워 스펙트럼에 대한 것이므로 식 (2.18)과 같은 코사인변환이 된다. 여기서 $y^{(m)}(k)$는 MFCC값들을 나타내고, $Y(m)$는 필터뱅크에서 얻어진 출력 값이며, M은 MFCC의 차수이다.

$$y^{(m)}(k) = \sum_{m=1}^{M} \log\{|Y(m)|\} \cos\left(k\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right) \qquad k = 0,1,2,\ldots,L \tag{2.18}$$

식 (2.18)에서 0차 MFCC계수 $y_t^{(0)}(k)$는 해당 프레임에서의 로그 에너지와 유사하며, 이때 이 계수는 시간 영역에서 쉽게 구해질 수 있다.

(7) Delta-MFCC

정적인 특징벡터는 임의의 시간에서 그 순간의 특성만을 나타낼 뿐 인접 벡터들 간의 관계를 고려하지 못한다. 이런 특성을 반영한 특징벡터인 Delta-MFCC는 식 (2.19)를 사용하여 구한다[17].

$$\Delta C_m(t) = \frac{\sum_{n=-K}^{K} n\, w(n)\, C_m(t+n)}{\sum_{n=-K}^{K} n^2\, w(n)} \tag{2.19}$$

여기서, $C_m(i)$는 i번째 프레임의 m번째 특징벡터 계수를 나타낸다. $w(n)$은 사각형의 윈도우 함수이며, 본 연구에서는 K는 2로 사용하여 앞뒤 두 프레임의 변화량을 고려하였다.

제 3 장 DTW를 이용한 단어 인식

추출된 특징벡터를 이용하여 음성을 인식하는 여러 가지 방법들이 연구되었는데, 대표적인 방법으로 DTW(dynamic time warping), HMM(hidden markov model), 신경망 모델(neural network) 등이 있다.

DTW는 동적 프로그램(dynamic programing) 알고리즘[18]을 이용하여 기준이 되는 신호의 패턴과 입력된 음성신호를 비교한 후, 가장 유사도가 높은 것을 찾아서 인식하는 방법이다[19, 20].

HMM은 조음 기관의 상태에 따른 음성신호의 변화를 통계적으로 모델링하는 방법으로, 1980년대 후반부터 성행된 음성인식 기술의 알고리즘이다. 이 방법은 음성신호의 변동을 확률적으로 취급하므로, 개인차나 조음 결합 등의 영향으로 나타나는 음성 패턴의 변화를 보다 정확하게 반영 할 수 있다. 그러나 모델의 구조를 결정할 때, 시행착오나 학습에 의한 경우가 많고, 학습시 대량의 데이터를 필요로하는 단점이 있다[5].

신경망은 HMM과 같이 음성신호의 변화를 망의 가중치로 변환해서 음성을 인식하는 방법이다. 이 방법은 개인차나 음성신호의 길이에 따른 변화를 망의 가중치에 적용하기가 어렵다. 신경망 모델을 음성인식에 적용시킨 방법으로 LVQ(learning vector quantization), TDNN(time delay neural network) 등이 있다[8].

본 연구에서는 단어 수량이 많지 않을 뿐 아니라 학습데이터가 충분하지 않은 경우에도 쉽게 이용할 수 있는 DTW를 이용하여 조타명령의 단어단위 인식기로 사용하였다. 본 장에서는 먼저 DTW 알고리즘을 살펴본 후, DTW를 이용한 단어 인식에 대하여 설명한다.

3.1 DTW

음성신호는 발화되는 내용뿐만 아니라 발화자의 서로 다른 특성 때문에 많은 다양성을 가지고 있으며, 이러한 개인적 특성도 화자의 상태 등에 따라 매우 유동적으로 변화한다. 그림 3.1은 “starboard”라는 동일한 단어에 대해 나타나는 서로 다른 모양의 파형을 도시한 것으로, 파형 A는 여성화자가, B와 C는 동일한 남성화자가 발화한 것이다. 이 그림에서 볼 수 있듯이 다른 화자들 간은 물론 동일화자라도 파형이 상이함으로, 이들 음성신호로부터 추출한 특징벡터들을 직접적으로 비교하는 방법을 통한 인식에서는 오인될 확률이 매우 높다.

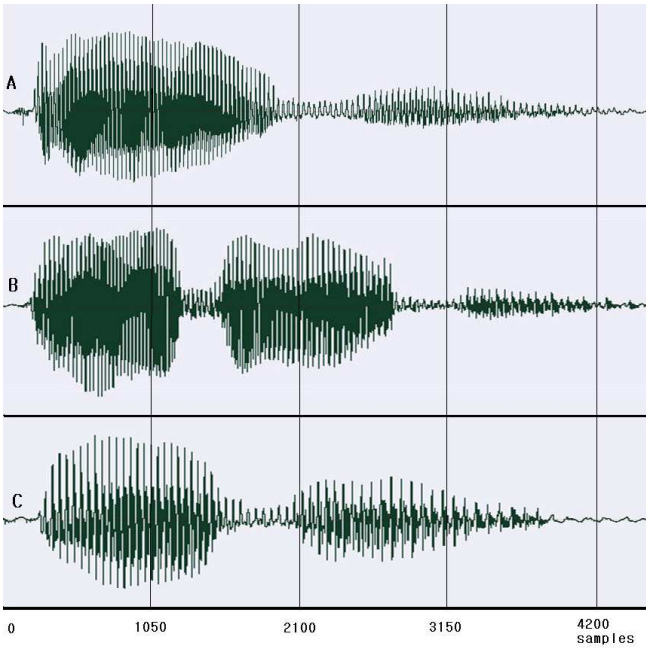


그림 3.1 “starboard” 단어의 음성신호.
Fig. 3.1 Different voice signal of “starboard”.

동적 프로그래밍 정합을 이용한 DTW 알고리즘은 비교 패턴들 간의 시간신축을 통하여 패턴 간의 비교 구간을 재정렬한 후, 유사도를 측정하므로 직접 비교를 통한 인식 방법보다 더 나은 인식 결과를 얻을 수 있다. DTW 알고리즘은 동적 프로그래밍 정합과정에서 많은 계산이 요구되어 인식 시간이 많이 걸리는 단점이 있다. 구현이 용이한 반면 많은 계산량이 요구되는 DTW 알고리즘의 이러한 특성 때문에, DTW 알고리즘은 주로 소용량 고립 단어 인식에 이용되고 있다[5, 19]. 아래에 DTW 알고리즘에 대해 설명한다.

(1) DTW 알고리즘

DTW 알고리즘을 이용한 인식 방법은 패턴 정합(pattern matching)방법의 하나로, 발음 속도의 변화를 동기화시켜 두 음성신호의 특징벡터들에 대한 비교를 가능하게 하는 것이다.

DTW 알고리즘을 이용한 패턴인식의 첫 단계는 비교할 두 개의 음성신호 사이의 거리를 구하는 것으로, 거리 계산에는 이들로부터 추출한 프레임별 특징벡터를 사용하며, 특징벡터는 식 (3.1)과 같다[5, 18].

$$\begin{aligned} S_x &: (X_1, X_2, X_3, \ldots, X_{T_x}) \\ S_y &: (Y_1, Y_2, Y_3, \ldots, Y_{T_y}) \end{aligned} \tag{3.1}$$

여기서 X_i, Y_j는 각 음성신호의 i와 j번째 프레임의 특징벡터이다. 두 음성신호 사이의 거리 $D(S_x, S_y)$는 식 (3.2)와 같이 프레임간 거리 $d(X_i, Y_j)$에 대한 함수로 정의되며, warping함수라고도 한다.

$$D(S_x, S_y) = F(d(X_i, Y_j)), \quad i=1,2,\cdots, T_X \text{ and } j=1,2,\cdots, T_Y \tag{3.2}$$

프레임간의 거리 $d(X_i, Y_j)$를 구하기 위해 다양한 거리측정 방법이 사용될 수 있지만, 본 연구에서는 MFCC 특징벡터에 대하여 가장 좋은 인식 결과를 보여주는 유클리디언 거리(euclidean distance) 방법을 이용하여 프레임간의 거리를 구한다[21]. 그림 3.1은 식 (3.3)을 사용하여 서로 다른 화자의 “starboard” 발성에 대한 프레임별 특징벡터의 유클리디언 거리를 나타낸 것이다.

$$d(X_i, Y_j) = \sqrt{\sum_{k=1}^{K} \left(C_k^{X_i} - C_k^{Y_j} \right)^2} \tag{3.3}$$

여기서 K는 MFCC 특징벡터의 차원이다.

패턴인식에 있어 시간적으로 동떨어져 있는 프레임 사이의 거리를 측정하는 것은 무의미하다. DTW 알고리즘은 이러한 점을 고려하여 여러 가지 제약조건들을 적용하여 불필요한 프레임간의 거리 측정을 생략한다. 즉, 한 음성의 모든 프레임을 각각 다른 음성의 모든 프레임과 비교하는 것이 아니라 대응되는 프레임들만 비교하여 계산량을 줄인다. 식 (3.2)의 함수 $F(d(X_i, Y_j))$는 가능한 두 프레임간 거리 중에서 몇 개만을 선택하여 전체적인 거리를 계산하기 위한 제약조건들로 다음절에서 설명한다. 여러 가지 제약조건에 의해 대응된 프레임간의 거리를 측정하여 최소의 거리를 갖는 프레임들 사이를 연결하면 그림 3.1과 같은 하나의 선이 만들어 지는데 이를 warping 경로라 한다.

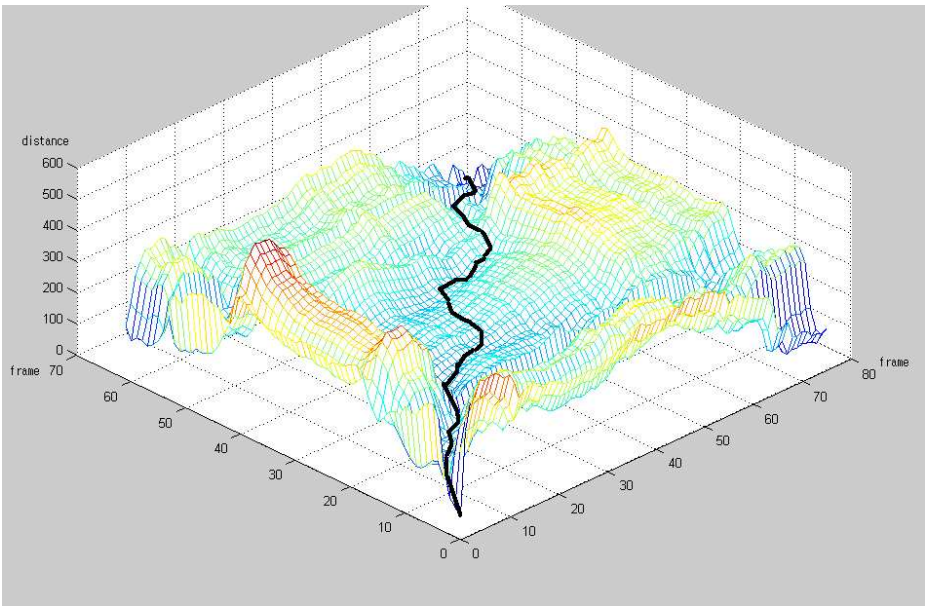


그림 3.1 두 특징벡터로부터 구한 프레임간의 유클리디언 거리.
Fig. 3.1 Euclidean distance between two MFCC features.

(2) Warping 함수의 제약 조건

식 (3.2)의 warping 함수 $F(d(X_i, Y_j))$는 음성 패턴의 시간축 변화를 나타낸 것으로 시간축 변화 성질을 따르게 된다. 본질적으로 음성 패턴의 시간축 구조는 연속성, 단조성, 음향적인 요소, 전이 속도의 제한 등을 고려해야 하기 때문에 음성 인식에 적용되는 DTW 알고리즘은 다음과 같은 5가지 제약 조건하에 이루어진다[19].

첫째, 끝점 제약(endpoint constraints) 조건은 같은 단어에 대한 발화 시간은 같다는 가정으로, 입력 패턴 $x(n)$과 참조 패턴 $y(m)$의 시작점과 끝점을 일치시킨다. 즉, 길이가 다른 두 음성신호도 DTW 과정을 통한 시간신축의 결과로 길이가 같아지는 효과를 가지게 되므로 끝점 제약 조건을 적용한다.

둘째, 단조 조건(monotonic condition)은 최적 경로가 항상 단조 증가해야만 한다는 것이다. 그림 3.2에서 입력 패턴과 참조 패턴의 최적 경로가 (i_k, j_k)에서 (i_{k+1}, j_{k+1})로 증가될 때 식 (3.4)와 같이 각각의 입력 패턴과 참조 패턴도 함께 증가된다는 것을 말한다. 음성신호에서 구한 특징벡터들은 언어적 의미를 내재하고 있고, 또한 시간의 흐름을 가지고 있기 때문에 알고리즘을 수행할 때 단조 증가 조건을 이용하는 것이 바람직하다.

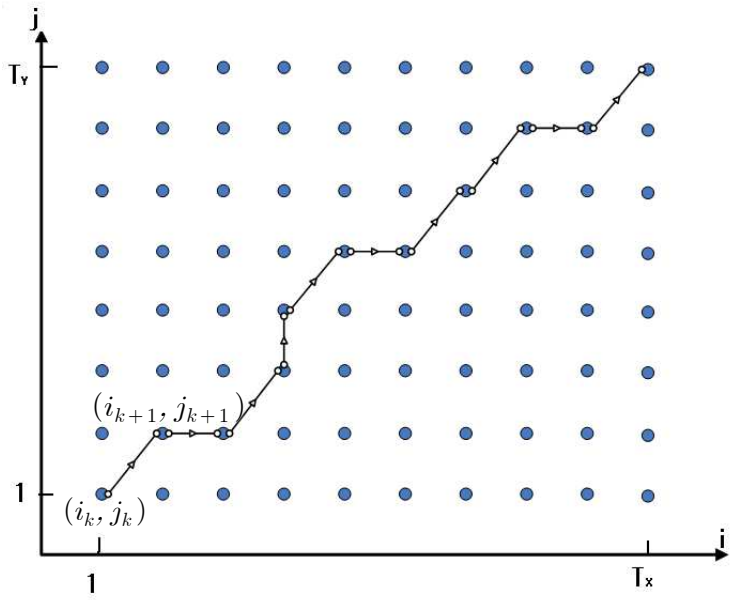


그림 3.2 Warping 경로.
Fig. 3.2 Warping path.

$$i_{k+1} \geq i_k, \qquad j_{k+1} \geq j_k \tag{3.4}$$

셋째, 국부 경로 제한 조건(local path constraint condition)은 격자 상의 한 노드에 도달하기 위한 경로에 제한을 둠으로써 시간축의 과다한 압축이나 팽창을 방지하는 역할을 한다.

넷째, 전역 경로 제한 조건(global path constraint condition)은 입력 음성 패턴과 참조 패턴간의 전 구간에 걸쳐 허용 가능한 영역을 제한시킴으로써, 인식시 오류와 계산량을 줄여 준다. 전역 경로 제한 조건은 국부 경로 제한 조건 때문에 최적 경로가 진행할 수 있는 평면이 제한되는 것과 동일 음성이 발성될 때 지속 시간의 차이가 1/2에서 2배를 넘지 않는다는 가정을 전제로 한다.

본 논문에서는 그림 3.3과 같은 전역 경로 제한을 적용하였는데, 그림에서 번호 ❶, ❷, ❸, ❹는 식 (3.5)에 나타낸 경로제한 함수를 적용한 결과이다.

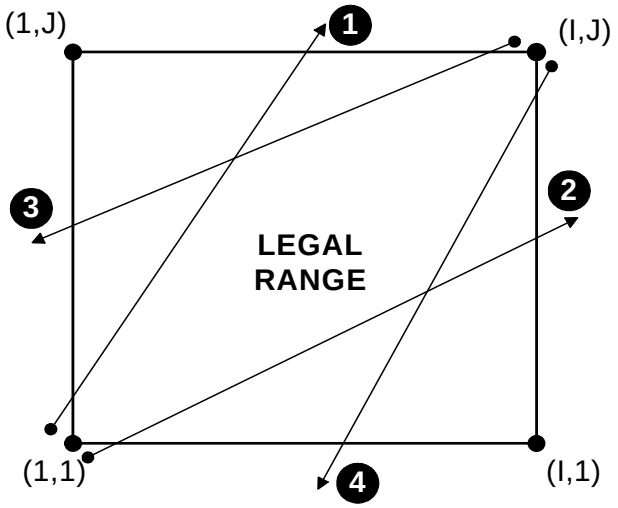


그림 3.3 전역 경로 제한 조건.
Fig. 3.3 Global path constraint condition.

$$s = \begin{cases} 2, & I = J \\ \frac{I}{J}, & otherwise \end{cases} \quad (3.5)$$

❶ $(i+1)\times s_1 + 3, \quad s_1 = s \times 2.0,\ s_2 = s \times 0.5$
❷ $(i-1)\times s_2 - 3$
❸ $(i-J)\times s_1 + I + 3$
❹ $(i-J)\times s_2 + I - 3$

다섯째, 기울기 가중치(slope weighting) 조건은 국부 경로의 가중치를 계산할 때, 모두 동일한 가중치를 주지 않고 기울기에 따라서 서로 다른 가중치를 둠으로써, 시간에 대해 불연속적으로 변하는 것을 방지하는 역할을 한다.

그림 3.4는 본 논문에서 제안한 시스템에 적용한 국부 경로 제약 조건을 나타낸 그림으로 3방향으로만 이동이 가능하도록 가중치를 부여하였다. 그림 3.4에서 $D(i_k, j_k)$는 그림 3.2와 같이 기준 패턴과 입력 패턴 프레임간의 관계를 격자 구조 나타내었을 때, 시작 지점에서 (i_k, j_k) 지점까지의 누적된 거리이며, $d(i_k, j_k)$는 (i_k, j_k)에서의 프레임간 거리이다.

그림 3.4는 ②의 이동 방향에 대하여, ①과 ③의 이동 방향보다 2배 큰 가중치를 부여한 것이다. 그림 3.4의 ①, ②, ③의 이동 방향 중 최적의 경로는 식 (3.6)을 이용하여 구한다.

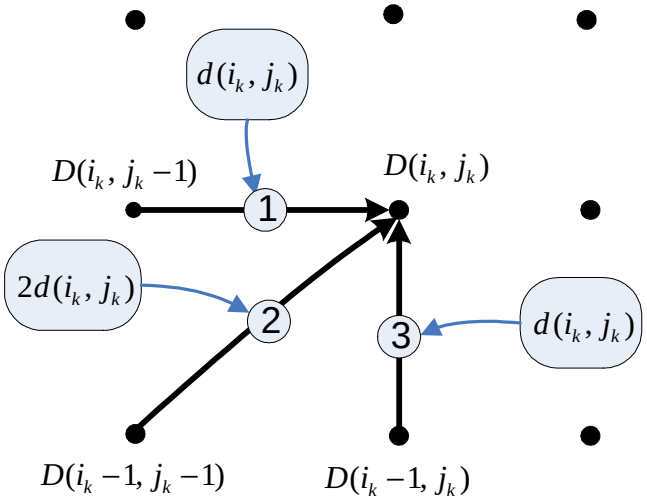


그림 3.4 지역 제한 조건과 기울기 가중치.
Fig. 3.4 Local constraints and slope weights.

$$\begin{aligned} &① \quad D(i_k, j_k-1) + d(i_k, j_k) \\ &② \quad D(i_k-1, j_k-1) + 2d(i_k, j_k) \\ &③ \quad D(i_k-1, j_k) + d(i_k, j_k) \end{aligned} \tag{3.6}$$

3.2 DTW를 이용한 단어 인식

선박의 조타명령은 23개의 제한된 단어들의 조합으로 구성되며 규칙적인 명령 형식을 따라 발화된다. 본 연구에서 제안한 음성 조타명령 인식기는 소용량의 고립단어 인식에 적합한 DTW 알고리즘의 특성을 효과적으로 활용하기 위하여 인식단위를 조타명령을 구성하는 단어단위로 설정하였다.

제안한 음성 조타명령 인식기는 23개 단어들의 조합으로 이루어진 임의의 조타명령이 입력되면 음성구간을 검출하고, 검출된 각각의 음성구간들을 인식할 단어단위의 입력 패턴으로 한다. 검출된 개별 단어의 신호들로부터 프레임별

특징벡터를 추출하여, 기준 패턴과 함께 DTW 알고리즘을 이용한 인식기에 적용시켜 얻어진 유사도에 따라 단어단위의 인식을 행한다.

제안한 조타명령 인식기는 미지의 입력 단어를 인식하기 위해, 먼저 비교의 기준이 되는 단어별 기준 패턴을 설정하여야 한다.

(1) 기준 패턴 생성

본 연구에서 제안한 인식기를 비롯한 여러 종류의 음성 인식기들이 사용하는 인식 기법들은 대부분 패턴 정합과정을 이용한다. 패턴 정합을 통한 인식 기법에 있어 인식률에 큰 영향을 미치는 요인 중의 하나가 기준 패턴이다. 한 클래스의 기준 패턴이 그 클래스의 특징을 충분히 대표하지 못하는 경우, 입력 패턴을 오인하는 경우가 많아져 인식률이 떨어진다.

제안한 인식기에서는 기준 패턴의 대표성을 잘 반영하기 위하여, 실제 인식에 사용하는 DTW 알고리즘을 이용한 방법을 똑 같이 적용하여 다음과 같은 과정으로 기준 패턴을 구한다. 먼저, 식 (3.7)과 같이 단어별로 준비된 N개 입력 패턴들 중에 n번째를 선택하여 나머지 $N-1$개의 입력 패턴들과 DTW 알고리즘을 적용해 거리를 모두 구한다. 다음으로 앞에서 구한 거리를 가지고 평균 거리 $\overline{D}(S_i^n, S_i)$를 산출한다. 이러한 평균 거리 산출과정을 N개의 모든 입력 패턴들에 대하여 개별적으로 실시한다.

$$\overline{D}(S_i^n, S_i) = \frac{1}{N-1}\left\{ \sum_{m=1}^{N} D(S_i^n, S_i^m) \right\}, \quad n = 1, 2, \ldots, N \tag{3.7}$$

다음으로, 식 (3.8)과 같이 단어별로 구한 $\overline{D}(S_i^n, S_i)$들 중 최소를 찾아 그 때의 입력 패턴 S_i^m을 그 단어의 기준 패턴 $S_{ref(i)}$로 설정한다.

$$S_{ref(i)} = \arg\min_{n} \{ \overline{D}(S_i^n, S_i) \} \tag{3.8}$$

(2) 단어단위 인식

단어단위 인식의 첫 단계로 미지의 입력 패턴에 대하여 DTW 알고리즘을 적용하여 앞에서 구한 N개의 각 기준 패턴들과의 거리 $D(S_x, S_{ref(i)})$를 식 (3.2)와 같이 각각 구한다. 다음으로, 각 기준 패턴과 비교를 통하여 구한 N개 거리들의 크기를 조사하여, 거리가 최소인 경우의 기준 패턴인 $S_{ref(i)}$에 해당하는 단어로 입력 패턴을 인식한 것으로 분류한다. 이를 나타내면 식 (3.9)와 같다.

$$\hat{w} = \arg\min_{S_{ref(i)}} D(S_x, S_{ref(i)}), \quad i=1,2,\ldots,N \tag{3.9}$$

여기서 $\hat{w}$는 인식된 단어이다.

앞에서 언급한 단어단위의 인식 결과를 병합하는 것으로 조타명령 인식을 수행할 수 있다. 그러나 이 경우, 인식된 단어들 중에 조타명령을 구성하는 단어와 하나라도 틀린 단어가 포함되어 있으면 명령단위의 인식은 오인되어져, 조타명령의 인식률은 단어단위 이상의 인식률을 기대하기 어렵게 된다.

본 연구에서는 조타명령의 인식률을 높이기 위하여 조타명령의 형식이 규칙적인 점과 조타명령을 구성하는 23개 단어들 중 명령의 시작 단어로 올 수 있는 것은 7개 정도인 것에 주목하여, 이를 명령단위 인식에 적용한다. 즉, 조타명령에 대한 문법을 생성하여 단어단위 인식에 이은 명령단위 인식에 적용하고, 덧붙여 단어단위의 오인식을 명령단위 인식에서 보상하는 방법도 적용한다.

명령단위 인식에서 단어단위의 오인식을 보상하는 방법은 다음과 같다. 식

(3.9)에서 거리 $D(S_x, S_{ref(i)})$가 최소인 것만 찾는 것이 아니라, 거리가 작은 순서대로 최소인 것부터 차례로 L개의 단어단위 인식 후보들을 선택한 다음, 문법을 적용한 명령단위 인식에 활용할 수 있도록 한다. 본 논문에서 제안한 인식기에서는 실험에 의해 L을 5로 설정하였다.

다음 장에 조타명령의 인식률을 향상시키기 위하여 단어단위 인식 후보들에 적용할 조타명령 문법에 대하여 설명한다.

제 4 장 문법에 의한 조타명령 인식

조타명령의 명령단위 인식은 조타명령을 구성하는 단언단위 인식 결과로 얻은 단어별 후보 L개에 대하여 word lattice를 구성한 다음, 조타명령 형식의 규칙을 적용하여 생성한 문법을 통하여 인식을 수행한다. 본 장에서는 먼저, 조타명령 문법이 적용될 인식의 대상이 되는 조타명령의 종류를 살펴보고, 다음으로 생성한 조타명령 문법에 대하여 설명하고, 마지막으로 단어단위 인식에서 얻은 단어별 후보로 구성된 word lattice에 문법을 적용하여 조타명령을 인식하는 방법을 설명한다.

4.1 조타명령

인식대상이 되는 조타명령의 예와 의미를 표 4.1에 나타내었으며, 표 4.2는 표 4.1에 예를 든 조타명령들을 구성하는 23개 단어를 보여주고 있다. 표 4.1의 예에서 볼 수 있는 것처럼, 하나의 조타명령은 표 4.2에 있는 1개에서 4개까지의 단어들의 조합으로 만들어지며, 조합되는 형식이 규칙적임을 알 수 있다.

이러한 조타명령을 구성하는 단어들의 조합 규칙을 이용하여 명령단위 인식에 적용할 문법을 생성한다.

표 4.1 선박 조타명령의 예.

Table 4.1 Example of steering command.

동작 구분	명령어	실행 내용
Rudder만으로 이루어지는 조타	starboard(or port) easy	타각을 우현(또는 좌현)으로 7°~8° 주시오.
	starboard(or port) ten	타각을 우현(또는 좌현)으로 10° 주시오.
	starboard(or port) fifteen	타각을 우현(또는 좌현)으로 15° 주시오.
	starboard(or port) twenty	타각을 우현(또는 좌현)으로 20° 주시오.
	hard starboard (or port)	타각을 우현(또는 좌현)으로 30°~35°(최대) 주시오.
	midship	타를 중앙으로 하시오.
	easy to ten	타각을 많이 준 상태에서 10°로 줄이시오.
Rudder와 compass로 이루어지는 조타	steady three one five	침로를 315°로 정침하시오.
	steady zero six two	침로를 62°로 정침하시오.
	steady	현재 침로로 정침하시오.
	course again	원래 침로로 정침하시오.

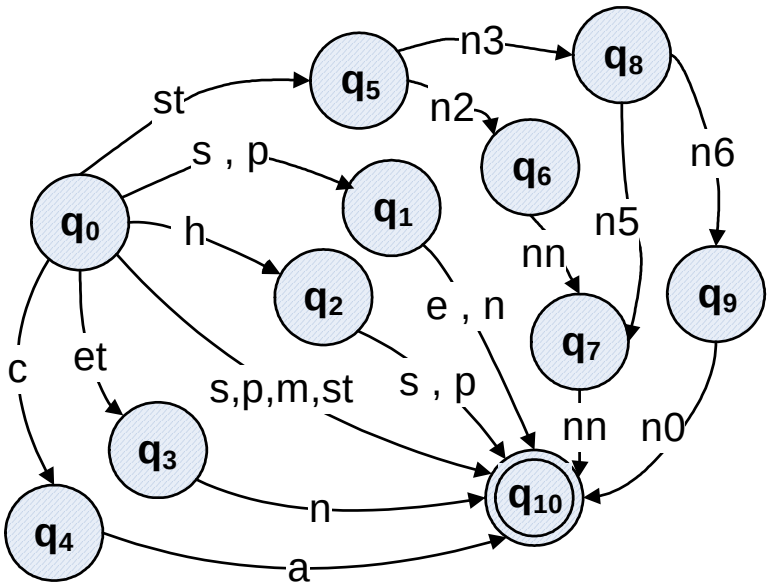
표 4.2 조타 음성명령을 구성하는 단어.

Table 4.2 Word of steering command.

word id.	약식 발음	발음할 명령어
1	/오/	0
2	/제로/	0
3	/원/	1
4	/투/	2
5	/쓰리/	3
6	/포/	4
7	/파이브/	5
8	/식스/	6
9	/세븐/	7
10	/에잇/	8
11	/나인/	9
12	/텐/	10
13	/피프틴/	15
14	/투엔티/	20
15	/스타보드/	starboard
16	/포트/	port
17	/하드/	hard
18	/미찝/	midship
19	/이지/	easy
20	/스테디/	steady
21	/코스/	course
22	/어게인/	again
23	/이지투/	easy to

4.2 조타명령 문법

조타명령의 규칙에 따른 상태 천이도는 그림 4.1과 같다. 다음 그림에서 q_0은 초기상태이며, 조타명령을 구성하는 23개의 단어들 중 7개만이 다음 상태로 천이가 가능한 입력이 될 수 있다. 모든 조타명령들은 입력에 따라 상태 천이가 일어난 후 종료상태인 q_{10}에서 끝나게 된다.



s : starboard	n0 : 0
e : easy	n2 : 0~2
p : port	n3 : 3
h : hard	n5 : 0~5
m : Midship	n6 : 6
et : easy to	n : 1~10,15,20
c : course	nn : 0~9
a : again	
st : steady	

그림 4.1 조타명령의 상태 천이도.

Fig. 4.1 State transition diagram of vessel's steering command.

조타명령의 상태 천이도에 나타난 것과 같이 조타명령을 구성하는 단어들의 수는 최소 1개에서 최대 4개까지이다. 그림 4.1의 조타명령 상태 천이도를 바탕으로 유한오토마타를 적용하여 문법을 생성하고, 이를 단어단위 인식에서 얻은 L개의 단어별 후보들로 구성한 word lattice[22]에 적용한다.

4.3 조타명령인식을 위한 문법적용

조타명령에 대한 문법을 L개의 단어별 후보들로 구성한 word lattice에 적용하여 최종적인 인식을 수행하는 과정은 다음과 같다.

① 단어단위 인식에서 구한 L개의 단어별 후보들을 조합하여 그림 4.2와 같이 임의의 조타명령을 구성한다.

② 구성된 word lattice에 조타명령 문법을 적용시켜, 문법 규칙에 맞는 경로만 선택한다.

③ 선택된 경로에 해당하는 명령들에 대한 전체 유사도를 각각 구한다.

④ 구하여진 경로별 유사도를 비교하여 가장 유사도가 큰 경로에 해당하는 명령을 최종적인 인식 명령으로 결정한다.

그림 4.2는 3개의 단어로 구성된 명령어가 입력된 경우 구성한 word lattice의 예이다. 조타명령 문법을 적용한 결과, 첫 번째 후보집단의 c_{11}, c_{14} 후보가 명령의 초기 상태에서 입력 가능한 단어이며, 두 번째 후보집단의 c_{22}, c_{23}, c_{24}만이 첫 번째 후보집단에서 선택된 단어 c_{11}, c_{14}에 이어 올 수 있는 단어이다. 마지막 후보집단에서는 c_{31}, c_{33}만이 앞에서 설정된 일련의 단어들 다음에 위치할 수 있는 단어이다.

그림 4.2에서 문법을 적용하여 구성 가능한 명령어의 수는 12개이다. 이 가능한 경로들에 대하여 단어단위 인식단계에서 구해 놓은 거리를 적용해 유사도를

측정하고, 그 결과에 따라 최종적인 조타명령 인식을 결정한다.

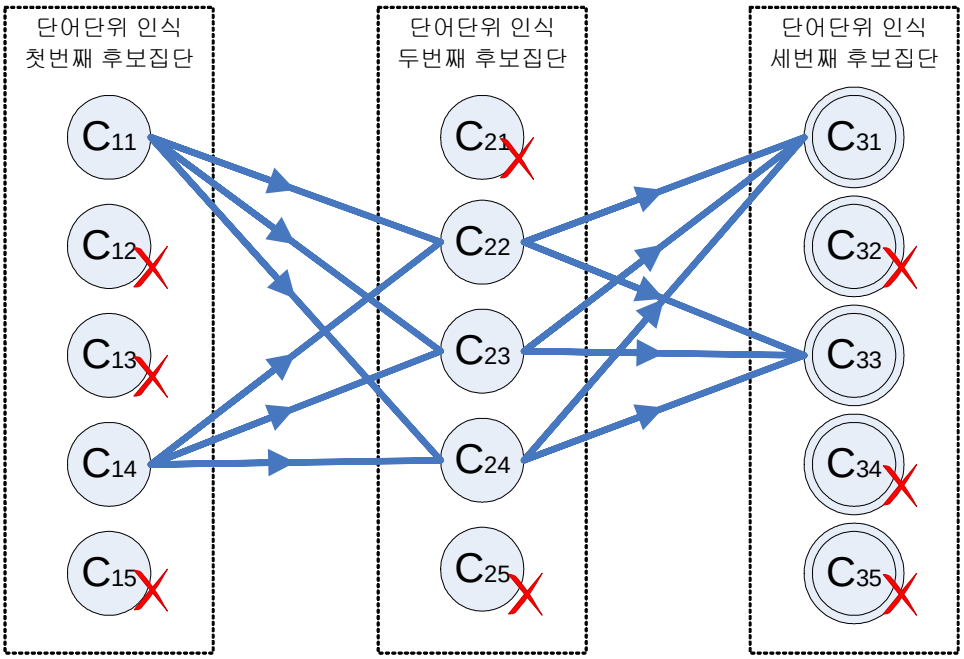


그림 4.2 단어단위 인식 후보집단을 이용한 조타명령 인식과정.

Fig. 4.2 State diagram of vessel's command recognition.

제 5 장 실험 평가 및 결과

본 논문에서 구성한 조타명령 인식 시스템의 성능을 알아보기 위하여 음성인식 실험을 다음과 같은 3가지 방법으로 수행하였다.

(a) 화자독립 단어 인식 실험

(b) 화자독립 명령 인식 실험

(c) 화자종속 명령 인식 실험

본 장에서는 실험환경과 음성데이터에 대해 먼저 설명한 다음 각각의 인식 실험과 결과에 대해 기술한다.

5.1 인식 실험 환경 및 음성 데이터

학습용 데이터와 시험용 데이터는 두 가지로 구성하였는데, 하나는 위의 실험 (a)와 (b)를 위한 것 (data set A)이고, 다른 하나는 실험 (c)를 위한 것 (data set B)이다. 전자는 33명의 화자 중 임의의 화자 23명(남 21, 여 2)을 선택해서 이들의 발화들을 모두 학습용 데이터로, 그리고 나머지 10명의 화자(남 9, 여 1)의 데이터는 시험용 데이터로 사용한 것이다. 후자의 경우에는 화자별로 세 번씩 발화된 단어별 데이터 중, 임의의 발화 하나를 화자별, 단어별 기준 발화로 사용하였으며, 나머지 두 개의 발화는 인식기의 평가를 위해 사용하였다.

시간 영역의 데이터를 2장에 기술한 것과 같은 방법으로 전처리, 특징벡터 추출과정을 통하여 MFCC를 추출하였는데 이 때 사용된 주요 파라미터들을 표 5.1에 보인다. 이렇게 추출된 MFCC 데이터로 3장에 기술한 DTW 기반 단어 인식 실험을 수행하고 이어서 4장에 기술한 word lattice를 사용하여 문법을 적용하여 최종적으로 조타명령 인식 결과를 얻었다. 인식기의 개발을 위해 Microsoft Visual C++ 6.0과 IBM 기종의 PC를 사용하였다.

표 5.1 음성 데이터와 전처리 조건.

Table 5.1 Speech data preprocessing conditions.

구 분		내 용
음성데이터	Sampling Rate	16㎑
	Sample당 bit수	16bit
전처리 조건	Pre-emphasis	0.95
	Hamming Window size	30ms
	사용된 특징벡터	MFCC
	특징벡터 차수	13차

5.2 화자독립 단어 인식 실험

본 실험에서는 조타명령을 구성하는 단어들을 인식하기위해 두 가지의 인식 실험을 수행하였는데, 하나는 MFCC만을 특징벡터로 사용한 실험이고, 다른 하나는 MFCC와 Delta-MFCC를 특징벡터로 사용한 실험이다. 인식 실험 결과는 표 5.2에 나타내었다. 이 표에서는 MFCC와 Delta-MFCC를 특징벡터로 사용한 실험에서의 인식률이 MFCC만을 특징벡터로 사용한 실험의 인식률보다 낮음을 알 수 있다. 이는 시간에 따른 특징벡터의 변화량인 Delta-MFCC가 시간 신축을 통하여 음성을 인식하는 DTW와 상충되어 나타난 결과로 보인다.

표 5.2 특징벡터 종류별 조타명령 단어 인식율.

Table 5.2 Recognition rates of two different feature vectors.

Word id.	MFCC (%)	MFCC + Delta-MFCC (%)
1(oh)	96.6	83.3
2(zero)	90	90
3(one)	80	93.3
4(two)	80	80
5(three)	56.6	60
6(four)	20	26.6
7(five)	80	80
8(six)	96.6	96.6
9(seven)	86.6	86.6
10(eight)	83.3	83.3
11(nine)	93.3	93.3
12(ten)	83.3	86.6
13(fifteen)	83.3	80
14(twenty)	70	70
15(starboard)	46.6	33.3
16(port)	90	90
17(hard)	73.3	73.3
18(midship)	53.3	53.3
19(easy)	30	36.6
20(steady)	46.6	43.3
21(course)	76.6	50
22(again)	50	36.6
23(easy to)	100	100
총 평균	72.7	70.7

5.3 화자독립 조타명령 인식 실험

항해에 사용되는 조타명령은 앞서 설명한 23개의 단어들의 조합으로 구성되

며, 조타명령이 될 수 있는 조합은 599가지이다.

이 실험에서는 단어단위 인식 실험에서 사용한 단어 음성데이터를 조합하여 가능한 조타명령들을 화자별로 만들어 인식 실험에 사용하였다. 표 5.3에 인식 결과를 나타내었는데, 문법을 적용했을 때가 적용하지 않았을 때보다 전체 인식률이 3%정도 증가되었다.

표 5.3 Word lattice를 사용한 명령단위 인식 실험.
Table 5.3 Result of command level recognition experiments.

실험 방법	인식률(%)
word lattice 미사용	67
word lattice+문법사용	70

표 5.4는 표 5.3의 첫 열에 보인 결과중에서 오인식을 분석한 것이다. 이들 중 특정한 몇 개의 단어에서 인식률이 아주 낮았는데, 그 이유는 다음과 같이 분석된다. 먼저 “four”의 경우 화자에 따라 무성음 /f/를 약하고 짧게 발음하여 “oh”로 오인되는 경우가 많았다. “easy”의 경우 발음상 같은 “easy-to”와 끝부분 발음이 비슷한 “three”로 많이 오인되며, 끝으로 “starboard”는 /s/발음의 간략화 현상에 의한 “hard”로 오인, 후반부 발음이 유사한 “port”로의 오인이 인식률을 저하시키는 것으로 나타났다.

표 5.4 MFCC 특징벡터를 사용한 단어별 인식 실험 결과의 오인식 분석.

Table 5.4 Analysis of mis-recognition in word-level recognition experiment.

word id.	MFCC (%)	잘못 인식한 단어 목록
1(oh)	96.6	four
2(zero)	90	four, two, seven
3(one)	80	port, five, nine
4(two)	80	seven, port
5(three)	56.6	eight, seven, six
6(four)	20	oh
7(five)	80	hard, ten
8(six)	96.6	three
9(seven)	86.6	ten, eight
10(eight)	83.3	three
11(nine)	93.3	five, seven
12(ten)	83.3	two, seven
13(fifteen)	83.3	zero, three
14(twenty)	70	eight, three
15(starboard)	46.6	hard, port
16(port)	90	oh, two, seven
17(hard)	73.3	five
18(midship)	53.3	easy to
19(easy)	30	easy to, three
20(steady)	46.6	ten, eight
21(course)	76.6	port, hard
22(again)	50	ten, five
23(easy to)	100	.
총 평균	72.7	

5.4 화자종속 조타명령 인식 실험

선박에서 조타명령은 항해사를 비롯한 소수의 인원만이 사용하는 것이다. 본 실험에서는 조타명령의 이러한 특성을 인식기에 반영하여, 화자종속 조타명령 인식기를 구성한 다음, 조타명령 데이터를 이용하여 인식 실험을 수행하였다.

이러한 조타명령의 특성을 인식기에 반영하여 화자종속 인식기를 구성하면 인식률을 높일 수 있을 것이다. 본 절에서는 화자종속 명령인식 실험 결과를 기술 한다.

표 5.5는 화자종속 조타명령에 대한 인식 실험 결과이다. 그림 4.1의 상태천이도에 따라 조타명령을 구성하는 첫 단어들을 기준으로 조타명령들을 구분하여 인식률을 구하였다.

표 5.5에 나타난 결과를 분석해보면 “course again” 명령이 다른 명령들에 비하여 오인된 경우가 많다. 그 이유를 살펴보면 다음과 같이 추정할 수 있다. 표 5.5의 단어별 인식 결과에서 “course”와 “again”은 타 단어들에 비해 낮은 인식 결과인 76.6%와 50%의 인식률을 보이고 있다. 조타명령 “course again”은 이 두개 단어들이 조합된 명령으로, 화자종속 인식의 경우에도 개별 단어의 인식 정도가 반영되어 나타난 결과로 볼 수 있다.

화자 종속적 음성 조타명령의 인식 실험 결과가 화자독립 인식에 비하여 인식률이 크게 증가되었음을 보여주는데, 이를 통해 제안한 화자종속적인 방법이 조타명령 인식기에 적절한 방법임을 알 수 있다.

표 5.5 화자종속 조타명령 인식률.

Table 5.5 Recognition rate of speaker dependent recognition experiment.

조타명령	인식률(%)	잘못 인식한 명령	비고
easy to <n>	97.9	hard port	n:1～10,15,20
course again	81.8	port 8, port 10	
hard <starboard | port>	98.4	hard port	
midship	98.4	steady	
port	100	·	
starboard	100	·	
steady	98.4	port	
starboard <easy | n>	97.8	hard port	n:1～10,15,20
port <easy | n>	97.6	hard port, hard	n:1～10,15,20
steady <000～360>	97.2	oh, four	
평균	97.2	·	

제 6 장 결론 및 향후 과제

선박의 조타음성명령을 인식하기 위해 본 논문에서 제안한 음성인식기는 조타명령의 특성에 맞추어, 소용량 고립단어 인식에 이용되는 DTW를 사용하였으며, 특징 추출 전단계로 인식률에 많은 영향을 미치는 음성끝점 검출의 정확도를 높이기 위해서 영교차율과 가중 엔트로피법을 적용하였다. 인식기의 후반부에는 조타명령에 대한 문법을 이용한 word lattice와 명령 구성 단어별 인식 후보를 적용하여 인식률 향상을 도모하였다.

음성조타명령 인식에 있어 실험 결과에서 볼 수 있듯이 화자종속 인식 방법이 적합했으며, DTW 알고리즘과 MFCC 특징벡터를 이용하여 효과적인 조타명령 인식이 가능함을 알 수 있었다.

제안한 시스템의 유효성을 검증하기 위하여 조타명령구성 단어별 및 조타명령별 인식 실험을 하였으며, 실험 결과 조타명령 구성 단어별 실험에서는 72.7%, 조타명령별 인식 실험에서는 화자독립의 경우 70%, 화자종속인 경우 97.2%를 보였다.

단어 인식의 경우, 몇 개의 단어가 인식률이 현격히 저하되어 나타났는데 이는 유사한 발음의 단어들에 의한 것으로 문법과 word lattice, 단어별 후보를 적용하여 보완할 수 있음을 결과를 통해 알 수 있었다.

제안한 인식기를 실제 선박에 적용시키기 위해서는 전처리과정에서의 잡음제거와 보안을 강화하기 위한 문장제시형 화자검증 등에 대한 지속적인 연구가 필요할 것이다.

참 고 문 헌

[1] 임영모, “산업판도를 바꿀 10대 미래기술,” 삼성경제연구소 CEO Information, No. 403, 2003.

[2] 서기열, “모형선박을 이용한 원격 조타제어시스템의 구축,” 한국퍼지및지능시스템학회 2003년 춘계 학술대회 학술발표 논문집, pp. 287-291, 2003.

[3] L. R. Rabiner, “Application of voice processing to telecommunications,” *Proc. of IEEE.*, Vol. 82, No. 2, pp. 199-228, 1994.

[4] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh Information Technology Series, 1990.

[5] L. R. Rabiner and B. H. Juang, *Fundamental of Speech Recognition*, Prentice-Hall Inc, 1993.

[6] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed., Springer-Verlag, 1972.

[7] A. V. Oppenheim and D. H. Johnson, “Discrete representation of signals,” *Proc. of ICASSP.*, pp. 681-691, 1972.

[8] D. R. Reddy, “Speech recognition by machine : a review,” *Proc. of IEEE.*, Vol. 64, pp. 501-531, 1976.

[9] M. G. Robert, *Entropy and Information Theory*, Stanford University, Springer-Verlag, New York, 1990.

[10] A. Ganapathiraju, L. Webster, J. Trimble, J. Bush and J. Kornman, “Comparison of energy-based endpoint detection for speech signal processing,” *Proc. of IEEE.,* Southeastcon, Tampa, Florida, USA, pp. 500-503, 1996.

[11] J. L. Shen, J. W. Hung and L. S. Lee, “Robust entropy based endpoint

detection for speech recognition in noisy environments," *Proc. of ICSLP.*, Sydney, Vol. 98, pp. 232-235, 1998.

[12] O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front-End for speech recognition in a noisy environment," Journal of phonetics, Vol. 16, pp. 109-123, 1988.

[13] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall International, 1993.

[14] S. B. Davis and P. Mermelstern, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentence," *IEEE Trans. on ASSP.*, Vol. 4, No. 28, pp. 357-366, 1980.

[15] B. H. Juang, L. R. Rabiner and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. on ASSP.*, 1987.

[16] E. Zwicker, "Subdivision of audible frequency range into critical bands," *JASA.*, Vol. 33, 1961.

[17] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on ASSP.*, Vol. 2, No. 29, pp. 254-272, 1981.

[18] S. Hiroaki and C. Seibi, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on ASSP.,* Vol. 26, No. 1, pp. 43-49, 1978.

[19] C. S. Myers and L. R. Rabiner, "A Level Building Dynamic Time Warping for Connected Word Recognition," *IEEE Trans. on ASSP.*, Vol. 29, No. 2, pp. 284-297, 1982.

[20] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, USA, 1957.

[21] P. Strobach, *Linear Prediction Theory*, Springer-Verlog, Berlin, Heidelberg, 1990.

[22] S. Saleem, S. C. Jou, S. Vogel and T. Schultz, “Using word lattice information for a tighter coupling in speech translation systems,” *ICSLP.*, Jeju Island, South Korea, 2004.

[23] S. S. Stevens and J. Volkmann, “The relation os pitch of frequency:A revised scale,” *Am. J. Psychol.*, Vol. 53, pp. 329-353, 1940.

[24] N. R. French and J. C. Steinberg, “Factors governing the intelligibility of speech sounds,” *JASA.*, Vol. 19, pp. 90-119, 1947.