Journal of the Research Institute of Industrial Technology. Vol. 23, 2006

DTW를 이용한 선박 음성조타명령 인식시스템의 구현

남언규* · 하동경* · 유강주* · 신옥근**
*한국해양대학교 대학원, **한국해양대학교 IT공학부 교수

Implementation of Vessel's Steering Command Recognition System using DTW

*E. K. Nam** · *D. K. Ha** · *G. J. Yu** *and O. K. Shin***
**Graduate school of National Korea Maritime University, Busan, Korea*
***Division of IT Engineering, National Korea Maritime University, Busan 606-791, Korea*

요 약 : 본 논문에서는 선박에서 항해사가 내리는 음성조타명령을 인식하는 시스템의 구현을 제안하고자 한다. 제안한 인식 시스템은 음성의 첫 부분과 끝부분을 찾아 얻은 음성신호에서 13차의 MFCC 특징벡터를 구하여 인식에 이용한다. 조타명령의 인식은 두 단계로 나누는데, 단어수준에서의 유사도 계산 단계와 가능한 명령들의 집합에서 조타명령을 선택하는 단계이다. 단어단위에서 입력신호의 MFCC 벡터와 기준 패턴들의 MFCC 벡터의 유사도는 DTW를 이용하여 계산한다. 다음으로 word lattice를 조타명령을 선택하는데 이용한다. 본 논문에서의 조타명령은 숫자가 포함된 23개의 단어로 이루어지며 1개에서 4개의 단어로 구성된다. 3가지의 인식실험으로 제안한 시스템의 성능을 평가하였는데, 첫 두 가지는 화자독립 단어인식 실험과 화자독립 명령어 인식 실험으로 약 70%의 인식률을 보여 효과적이지 못했다. 최종 실험은 화자종속 명령어 인식 실험으로, 실제 환경의 시스템에 적용해도 효율적일 수 있는 97% 이상의 인식률을 보였다.

핵심용어 : 선박 음성조타명령, 음성인식, DTW, word lattice, MFCC, Delta-MFCC.

ABSTRACT : *In this paper, we present an implementation of automatic speech recognition system to recognize vessel's steering command. The system detects the point at which the speech starts and ends, and then processes acoustic signals to produce the 13th order MFCC. Recognition of steering command is conducted in 2 steps: the calculation of similarity in word level, and 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 paper 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.*

KEY WORDS : vessel's steering command, DTW, speech recognition, word lattice, MFCC, Delta-MFCC.

1. 서 론

인간의 음성인식 과정을 자동화하고자 하는 자동음성인식(ASR:automatic speech recognition)에 대한 연구가 활발하게 진행되어 왔으며, 근래에는 산업 판도를 바꿀 미래의 10대 기술로 선정되기도 했다[1]. 이러한 음성인식 기술은 여러 가지 정보통신 산업 분야에 적용되어 실용화가 되어 왔으나, 해양 관련 산업 분야에 적용된 경우는 흔하지 않다[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 영교차율(zero crossing rate)과 가중 엔트로피(entropy)[3]법을 적용하였다. 인식기의 후반부에는 조타명령에 대한 문법을 이용

* sagent@hanafos.com, +82-51-410-4928
** okshin@hhu.ac.kr +82-51-410-4572

한 word lattice[4]와 명령 구성 단어별 인식후보를 적용하여 인식률 향상을 도모하였다.

본 논문에 사용된 음성데이터는 23개의 단어로 구성되며, 각각의 조타명령은 이들 단어를 1~4개 조합하여 만들어진다. 33명의 화자로부터 단어별 발화를 녹음하여 본 연구에서 구현한 시스템으로 인식 실험을 수행하였다. 그 결과, 화자독립형의 경우 70%, 화자종속형의 경우 97%의 인식률을 보여 화자종속 명령인식 방법이 효과적임을 알 수 있었다.

2. 음성인식 시스템의 구성

그림 1은 본 논문에서 제안한 음성 인식 시스템의 구조이다. 그림에서 "음성구간 추출"은 마이크를 통해 입력되는 음향신호로부터 음성구간을 추출하는 부분이다. 입력된 음성신호를 preemphasis, frameblocking, windowing 등의 처리를 시간영역에서 수행한 다음, 주파수영역으로 변환하기 위해 푸리에 변환(Fourier transform)을 수행한다[6]. 주파수 영역의 신호에 인간의 청각특성을 적용한 멜 필터 뱅크(Mel-filter bank)를 적용하여 MFCC 특징벡터를 계산하는 과정이 "특징벡터 추출"이다. 이렇게 얻은 입력신호의 특징벡터를 기준 패턴특징벡터와 DTW를 이용해서 유사도를 계산하고, 유사성이 높은 후보들을 찾는데 이 과정은 조타명령의 단어단위로 수행된다. 마지막으로 이상의 과정에서 얻은 단어 단위의 인식후보들에 조타명령 문법을 적용시켜 최종적인 조타명령을 인식하게 된다.

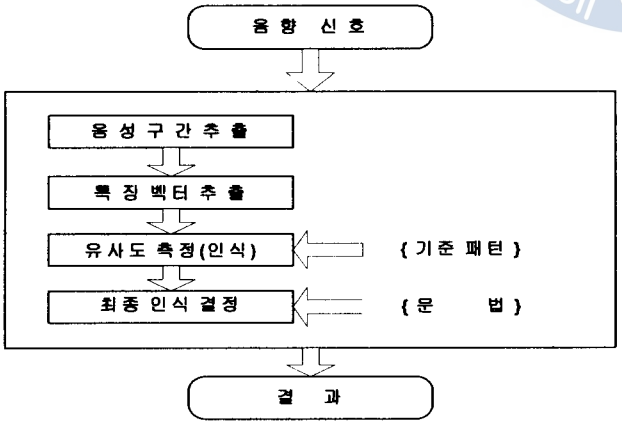


Fig. 1 The structure of the proposed speech recognition system.

2.1 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[6]. 본 논문에서는 영교차율과 엔트로피를 이용하여 음성구간을 추출하였다.

2.2 특징벡터 추출

특징벡터 추출은 시간 영역의 음성 신호로부터 음향학적 특징을 포함하는 벡터를 추출 하는 과정이며, 본 논문에서는 기존의 연구들에서 많이 활용되어 좋은 결과를 보이고 있는 MFCC와 delta-MFCC를 특징벡터로 사용하였으며, 이를 추출하는 과정은 그림 2와 같다.

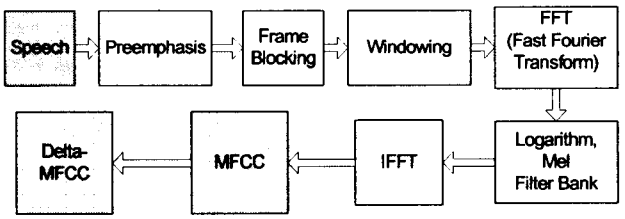


Fig. 2 Block diagram of MFCC and delta-MFCC extraction.

2.3 DTW를 이용한 단어 인식

DTW는 동적 프로그램(dynamic programing) 알고리즘[7]을 이용하여 기준이 되는 신호의 패턴과 입력된 음성신호를 비교한 후, 가장 유사도가 높은 것을 찾아서 인식하는 방법이다[8, 9]. 본 논문에서는 단어 수량이 많지 않을 뿐 아니라 학습데이터가 충분하지 않은 경우에도 쉽게 이용할 수 있는 DTW 방법을 음성조타명령의 단어단위 인식기로 사용하였다.

2.3.1 기준 패턴 생성

본 논문에서 제안된 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 (1)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 산출 과정을 N개의 모든 입력 패턴들에 대하여 개별

적으로 실시한다.

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

다음으로, 식 (2)와 같이 단어별로 구한 $\overline{D}(S_i^n, S_i)$들 중 최소를 찾아 그 때의 입력 패턴 S_i^n을 그 단어의 기준 패턴 $S_{ref(i)}$로 설정한다.

$$S_{ref(i)} = \min\{S_i^n : \overline{D}(S_i^n, S_i)\} \tag{2}$$

2.3.2 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)과 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 (4)와 같다.

$$D(S_x, S_y) = F(d(X_i, Y_j)) \quad i = 1,2,\cdots,T_X \quad \text{and} \quad j = 1,2,\cdots,T_Y \tag{3}$$

$$\hat{w} = \underset{S_{ref(i)}}{\arg\min}\, D(S_x, S_{ref(i)}) \tag{4}$$

여기서 T_X와 T_Y는 각 신호의 프레임 개수이며, $\hat{w}$는 인식된 단어이다.

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

본 논문에서는 조타명령의 인식률을 높이기 위하여 조타명령의 형식이 규칙적인 점과 조타명령을 구성하는 20여개 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.

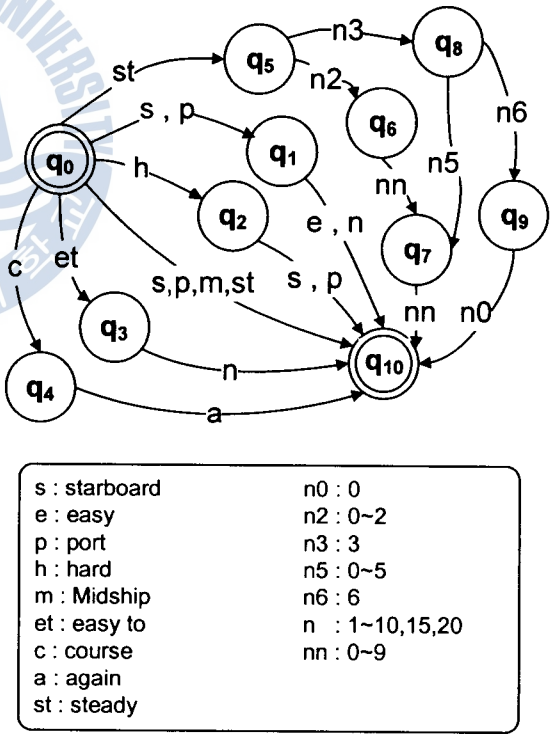
명령단위 인식에서 단어단위의 오인식을 보상하는 방법은 다음과 같다. 식 (4)에서 거리 $D(S_x, S_{ref(i)})$가 최소인 것만 찾는 것이 아니라, 거리가 작은 순서대로 최소인 것부터 차례로 L개의 단어단위 인식 후보들을 선택한 다음, 문법을 적용한 명령단위 인식에 활용할 수 있도록 한다.

3. 문법에 의한 조타명령 인식

조타명령의 명령단위 인식은 조타명령을 구성하는 단어단위 인식 결과로 얻은 단어별 후보 L개에 대하여 word lattice를 구성한 다음, 조타명령 형식의 규칙을 적용하여 생성한 문법을 통하여 인식을 수행한다.

3.1 조타명령 문법

조타명령의 규칙에 따른 상태 천이도는 그림 3과 같다. 다음 그림에서 q_0는 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	

Fig. 3 State transition diagram of vessel's steering command

조타명령의 상태 천이도에 나타난 것과 같이 조타명령을 구성하는 단어들의 수는 최소 1개에서 최대 4개까지이다. 그림 3의 조타명령 상태 천이도를 바탕으로 유한오토마타를 적용하여 문법을 생성하고, 이를 단어단위 인식에서 얻은 L개의 단어별 후보들로 구성한 word lattice에 적용한다.

3.2 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 4와 같이 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 4는 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 4에서 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.

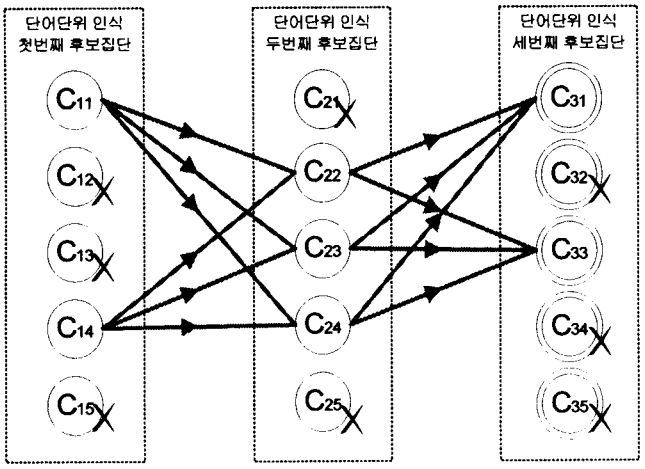


Fig. 4 State diagram of vessel's command recognition.

4. 실험 평가 및 결과

본 논문에서 구성한 음성조타명령 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

4.1 인식 실험 환경 및 음성 데이터

16㎑, 16bit로 샘플링(sampling)하여 수집한 음성데이터는 학습용 데이터와 시험용 데이터 두 가지로 분류하였는데, 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 데이터는 학습용과 시험용 화자를 따로 정하지 않고 화자별로 세 번씩 발화된 단어별 데이터 중, 임의의 하나의 발화를 화자별, 단어별 기준 발화로, 나머지 두 개의 발화를 시험용으로 이용하였다. 인식기의 개발을 위해 Microsoft Visual C++ 6.0과 IBM 기종의 PC를 사용하였다.

4.2 단어단위 인식 실험

조타명령을 구성하는 단어를 인식하기 위한 단어단위의 인식 실험에서는 MFCC만을 이용한 경우, MFCC와 MFCC의 시간 차원의 변화량인 delta-MFCC를 함께 이용하는 경우의 두 가지에 대해 인식 실험을 수행하여 표 1과 같은 결과를 얻었다.

인식 결과를 보면 일반적인 기대와는 달리 더 많은 정보와 계산량이 요구되는 후자의 특징 벡터를 적용한 경우 오히려 인식률이 저하되었는데, 이는 각 단어의 기준 패턴들이 한 개씩뿐인 것과 DTW가 시간신축을 통하여 유사도를 측정하는데, delta-MFCC가 시간축을 고려한 정보인 것이 상충되어 나타난 결과로 보여진다.

표1의 '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 보인다.

4.3 명령단위 인식 실험

항해에 사용되는 조타명령은 앞에서 실험한 23개의 단어들의 조합으로 구성된다. 문법을 적용하여 구한 가능한 조타명령은 599가지이다. 본 논문에서는 단어단위 인식 실험에서 사

Table 1 Recognition rates of two different feature vectors.

word id.	MFCC 13차 (%)	MFCC+ delta-MFCC 26차 (%)
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

용한 단어 음성데이터를 조합하여 가능한 조타명령들을 화자별로 생성하여 인식실험에 사용하였다.

제안한 음성조타명령 인식기에 적용되어질 조타명령은 실제 선박에서는 항해사를 비롯한 소수의 인원만이 사용하는 것으로, 이러한 조타명령 특성을 인식기에 반영하면 화자 독립적이 아닌 화자종속적인 특성을 가져야 한다. 즉, 소수의 특정인을 제외한 음성조타명령에 대해서는 거부를 함으로써 선박의 잘못된 운항을 사전에 방지할 수 있어야 한다. 제안한 인식기에 이러한 기능을 추가하기 위해서 그림 6과 같은 화자검증을 통한 화자 종속적인 인식 실험을 실시하였다.

5. 결론 및 향후 과제

선박의 음성조타명령을 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 명

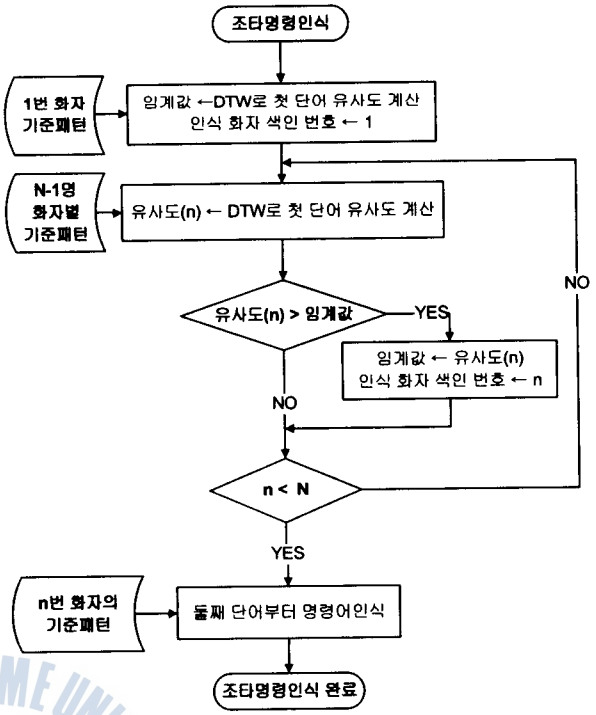


Fig. 6 Block diagram of a speaker identification.

령 구성 단어별 인식후보를 적용하여 인식률 향상을 도모하였다.

음성조타명령 인식에 있어 그림 7의 실험결과에서 볼 수 있듯이 화자검증을 통한 화자종속인식 방법이 적합했으며, DTW 알고리즘과 MFCC 특징벡터를 이용하여 효과적인 화자검증이 가능함을 알 수 있었다.

제안한 시스템의 유효성을 검증하기 위하여 조타명령 구성 단어별 및 조타명령별 인식 실험을 하였으며, 실험 결과 조타

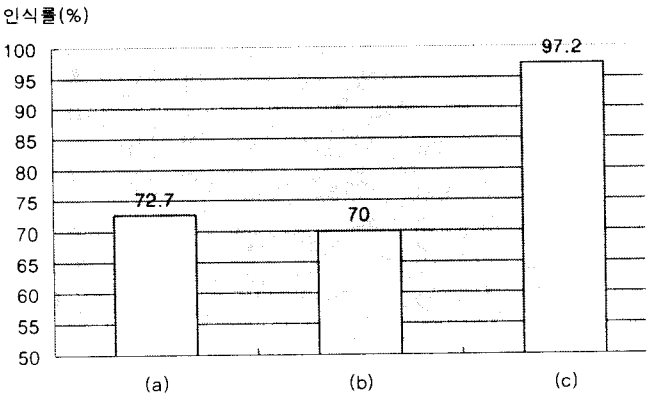


Fig. 7 The recognition rates of the experimental. (a)speaker independent, word recognition, (b)speaker independent, command recognition, (c)speaker dependent, command recognition.

명령 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 전처리 과정에서의 잡음 제거와 보안을 강화하기 위한 문장제시를 통한 화자검증 등에 대한 지속적인 연구가 필요할 것이다.

참 고 문 헌

[1] 임영모, "산업판도를 바꿀 10대 미래기술," 삼성경제연구소 *CEO Information*, No. 403, pp. 1-18, 2003.

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

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

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

원고접수일 : 2005년 12월 30일
원고채택일 : 2006년 1월 8일