1998년도 한국음향학회 학술발표대회 논문집 제 17권 제 2(s) 호

주행중인 자동차 환경에서의 고립단어 음성인식 연구

유봉근*, 이정기*, 김순협*, 박찬석**, 이순재**

* 광운대학교 컴퓨터공학과, ** 기아자동차 기술센터

A Study of Isolated Words Speech Recognition in a Running Automobile

Bong-Keun Yoo*, Jeong-Gi Lee*, Soon-Hyob Kim*, Chan Seok Park**, Soon Jae Lee**

* Kwangwoon University, ** KIA Motors

ybk@explore.kwangwoon.ac.kr

요약

본 논문은 주행중인 자동차 환경에서 운전자의 안전성 및 편의성의 동시 확보를 위하여, 보조적인 스위치 조작없이 음성의 입, 출력이 가능하도록 한다. 이때 잡음에 강인한 threshold 값을 구하기 위하여, 일정한 시간마다 기준 에너지와 영교차율(Zero Crossing Rate)을 변경하며, 밴드패스 필터(bandpass filter)를 이용하여 1차, 2차로 나누어 실시간 상태에서 자동으로, 정확하게 끝점검출(End Point Detection)을 처리한다. 기준패턴(reference pattern)은 DMS(Dynamic Multi-Section)를 사용하며, 화자의 변별력을 높이기 위하여 2개의 보너스를 제안한다. 또한 주행중인 차량의 잡음환경에 강인하기 위하여 일반주행(80km/h 이내), 고속주행(80km/h 이상)등으로 나누며 차량의 가변잡음 크기에 따라 자동으로 선택하도록 한다. 음성의 특징 백터와 인식 알고리즘은 PLP 13차와 One-Stage Dynamic Programming (OSDP)를 이용한다.

실험결과, 자주 사용되는 차량 편의장치 제어명령 33개에 대하여 충무, 영동 고속도로(사속 80km/h이상)에서 화자독립 89.75%, 화자종속 90.08%의 인식율을 구하였으며, 경부 고속도로에서는 화자독립 92.29%, 화자종속 92.42%의 인식율을 구하였다. 그리고 저속 주행중인 자동차 환경(80km/h 이내, 시멘트, 아스팔트등의 서울시내 및 시외도로)에서는 화자독립 92.89%, 화자종속 94.44% 인식율을 구하였다.

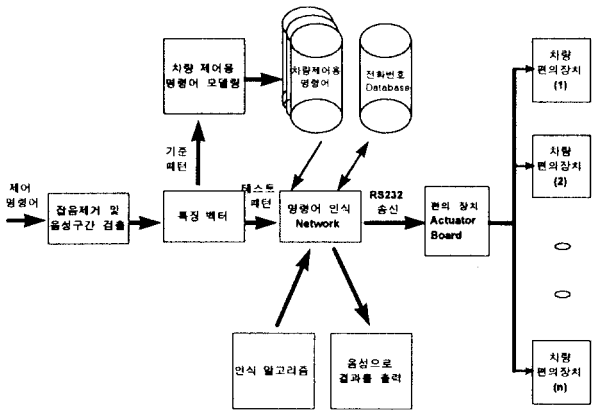
I. 서론

오늘날의 자동차 개발의 추세를 보면, 자동차의 기본 주행기능 이외에 다양한 편의장치 기능을 제공하게 됨에 따라 조작버튼의 수도 상대적으로 증가되고 있는 실정이다. 이 경우 차량의 운전시 동시에 각종 장치를 조작해야 함으로써, 운전에 대한 주의 및 집중력을 저하시켜 운전자의 안전운전에 상당한 악영향을 미칠 수 있다. 따라서 차량의 각종 편의장치를 조작을 음성으로 대신하게 된다면 주행시 운전자의 편의성과 안전성을 높일 뿐만 아니라, 자동차 문화의 부가적인 서비스를 제공하고, 자동차에 대한 인간의 친근감을 부여하게 될 것이다.

II. 본론

2.1. 시스템 개발 환경

음성입력은 핀-타입 전방향성 콘덴서 마이크를 통해 이루어지며, 11.025kHz 샘플링 주파수로 이산화되어 16bits로 양자화 된다. 이 과정은 일반적인 PC용 사운드 카드를 통해서 이루어지며 노트북 PC상에서 비주얼 C++를 이용하여 음성신호처리 및 알고리즘이 실시간으로 구현된다. 음성취득 및 실험은 서울시내, 시외도로, 공무, 영동, 경부고속도로에서 이루어지며 사용된 차량은 기아자동차의 포텐샤이다. 본 논문에서 사용한 단어는 비상등, 실내등, 에어컨등을 제어하기 위한 차량제어 명령어 22단어와 Voice Dialing을 위한 숫자음 11단어이며, 이들 단어는 주행중인 차량에서 화자와 마이크 거리를 약 30cm정도로 두고 취득하였다. 【그림 1】은 자동차 환경에서의 음성인식 시스템의 구성도를 보인다.



【그림 1】 자동차 환경에서의 음성인식 시스템

2.2. 음성구간 자동 검출

음성구간과 묵음구간을 분리해 내는데 가장 널리 이용되는 방법은 영교차율과 단구간 에너지(Energy)이다.[1] 하지만 주행중인 자동차의 배경잡음에서는 영교차율과 단구간 에너지만으로는 끝점검출을 한다는 것은 매우 어렵다. 따라서 본 논문에서는 영교차율과 단구간 에너지를 이용하기 전에 밴드패스 필터를 이용하여 잡음을 제거한다. 그리고 1차로 음성구간과 노이즈구간을 포함한 음성을 구하고, 다시 2차로 음성구간만을 구한다.

【그림 2】 음성구간 검출 전체 블록도

$$y(n) = \sum_{k=1}^{16} x(n-k)h(k) \quad (1)$$

식 (1)에서 $x(n)$은 배경잡음이 섞인 음성 신호이고, $h(k)$는 필터 계수이다. 이때 $y(n)$은 필터링한 출력값을 가르킨다.

$$ZCR(i) = \sum_{n=1}^{N-1} sgn(y_{n+(i-1)*N} \quad y_{n+1+(i-1)*N}) \quad (2)$$

$$E(i) = \sum_{n=1}^{N} |y((i-1)*N+n)| \quad (3)$$

$$sgn(y_n \times y_{n+1}) = \begin{cases} 1, & (y_n \times y_{n+1}) > 0 \\ 0, & (y_n \times y_{n+1}) < 0 \end{cases} \quad (4)$$

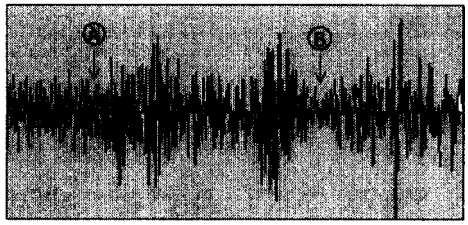
식 (2)와 (3)은 영 교차율과 단구간 에너지를 구하는 식으로 i 는 프레임의 번호이고 ZCR(i)는 i번째 프레임의 영교차율 수이다. N은 샘플수 128을 가르키며, E(i)는 i 번째 프레임의 에너지 값의 합을 의미한다. 식 (2)의 sgn은 식(4)와 같다. 본 논문에서는 음의 스키를 감소시킨 잡음과 음성구간을 좀 더 정확하게 구별하기 위하여 식 (2), (3)에 제곱을 하여 식 (5), (6)을 구하고, 기준 threshold는 일정한 시간마다 재조정하여 잡음에 강인하게 처리한다.

$$\overline{ZCR} = ZCR(i)^2 \quad (5)$$

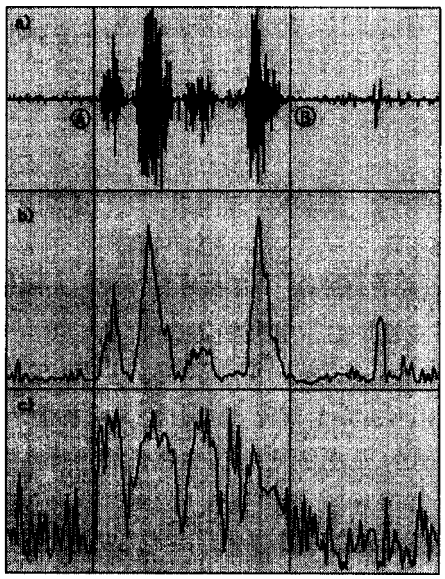
$$\overline{E} = E(i)^2 \quad (6)$$

2.3. 특징벡터

오프라인 실험을 수행한 결과 자동차 잡음환경에서는 PLP[2]가 LPC나 LPC 멜캡스트럼보다 인식율이 높은 것으로 나타났다.[3] 따라서 본 논문에서는 실제 주행중인 자동차 환경에서 특징벡터로 PLP를 사용한다.



【그림 3】 100km/h 전후의 속도에서 취득한 데이타

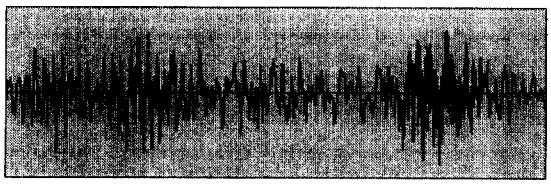


a) 【그림 3】의 잡음제거된 파형

b) a)신호에 대한 단구간 에너지

c) a)신호에 대한 ZCR

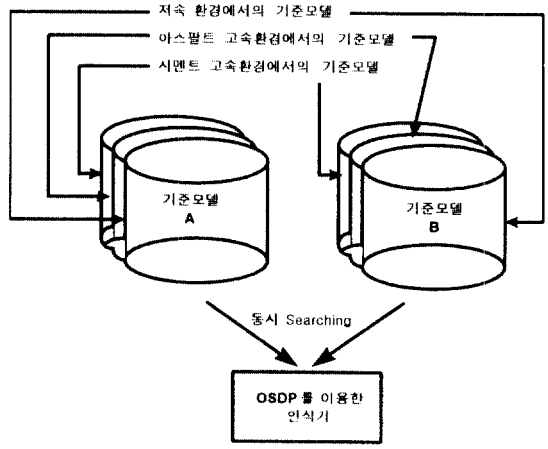
【그림 4】 잡음제거된 음성신호에 대한 영교차율과 단구간 에너지



【그림 5】 2차 끝점처리를 수행한 데이타

2.4. DMS 모델과 모델구성

DMS 모델은 유사한 특징을 가지는 벡터들을 한 구간으로 만들기 위하여, 구간을 동적으로 분할하여 특징벡터를 구함으로써 짧게 발음되는 특성까지도 대표 특징벡터로 선택될 수 있고 지속시간 정보도 갖도록 하는 장점을 가진 모델이다.[4] 즉, 단어 패턴의 연속된 대표값인 지속시간 정보와 대표특징 벡터를 사용하여 음성인식을 수행함으로써 파열음이나 파찰음 등과 같은 짧은 음소의 특징벡터도 동등한 비중을 가질 수 있도록 구간을 동적으로 나누어 준다.



【그림 6】 기준모델 구조

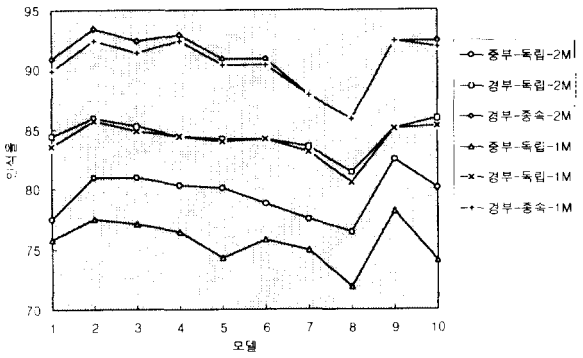
실제 시멘트 도로를 고속 주행중인 차량이라 할지라도 도로의 상황, 차속도, 주변환경에 따라 저속환경의 잡음이 발생할 수도 있고, 아스팔트 고속환경의 잡음이 발생할 수 있다. 따라서 본 논문은 【그림 6】과 같이 기준모델을 3개의 환경으로 나누었으며, 잡음크기에 따라 자동으로 선택하도록 하였다. 저속환경에서의 기준모델은 idle상태에서 취득한 데이터로 구성하며[3], 고속주행 환경에서는 40km/h에서 취득한 데이터와 idle상태에서 취득한 데이터로 구성한다.[5]

또한 화자의 변별력을 높이기 위하여 A, B 2개의 환경을 구성하였다. 【표 1】은 고속환경에서의 강인한 모델을 구성하기 위한 모델구조[5]이며, 【그림 7】은 A 모델만 사용하였을 경우와 A, B 모델 2가지를 동시에 기준모델로 사용했을 때의 인식율을 보이고 있다. 이 그래프에서 알 수 있듯이 1개의 기준모델을 사용하는 것보다 2개의 모델을 사용하는 것이 인식율이 높게 나옴을 알 수 있다. 그러나 단어수가 2배로 늘어남으로 처리하는데 속도가 다 걸리는 단점이 있다.

【그림 7】에서 중부, 경부는 중부(영동포함), 경부고속도로에서 취득한 데이터를 가르키며, 독립, 종속은 화자독립, 화자종속을 가르킨다. 또한 2M은 A, B 2개의 모델을 기준모델로 사용할 경우이며, 1M은 A 모델을 기준모델로 사용했을 경우를 의미한다.

【표 1】 모델 구성을 위한 환경

차속도 \ 모델		1	2	3	4	5	6	7	8	9	10
Idle	발음	2	1	2	3	1	3	3	3	0	0
40km/h	횟수	2	3	3	3	2	2	1	0	3	2



【그림 7】 A 모델과 A, B모델의 인식율

【표 2】 중부, 영동 고속도로의 인식율

화자 \ 횟수		1	2	3	4	5	6	계	인식률
화자독립	A 화자	93.94	93.94	96.97				94/99	89.75 (1007/1122)
		80-90km	80-90	75-85					
	B 화자	93.94	93.94	90.91	96.97	96.97	96.97	168/198	
		70-100km	85-100	70-95	80-110	70-100	95-105		
	C 화자	90.91	84.85	87.88	90.91	87.88	84.85	262/297	
		60-80km	80-100	90-100	75-90	85-100	80-90		
		87.88	87.88	90.91					
		70-90km	70-90	70-90					
	D 화자	90.91	90.91	90.91	90.91	87.88	90.91	179/198	
		70-100km	80-100	80-90	85-105	70-90	80-90		
	E 화자	90.91	84.85	93.94	90.91	87.88		148/165	
		80-90km	80-85	80-85	90-105	80-105			
	F 화자	84.85	81.82	81.82	78.79	84.85		136/165	
		70-90km	80-90	80-85	70-80	70-90			
화자종속	G 화자	78.79	90.91	87.88	90.91	84.85	87.88	235/264	90.08 (327/363)
		80-100km	80-90	80-100	70-100	80-100	85-105		
		96.97	93.94						
		80-100km	80-100						
	H 화자	90.91	93.94	93.94				92/99	
		80-110km	90-100	95-110					

【표 3】 경부 고속도로의 인식율

화자 \ 횟수		1	2	3	4	5	6	계	인식률
화자독립	A 화자	93.94	96.97	93.94				94/99	92.29 (335/363)
		75-95km	85-95	90-95					
	B 화자	93.94	96.97	93.94	96.97	96.97		158/165	
		80-110km	100-110	95-110	100-105	90-100			
	C	84.85	81.82	84.85				83/99	
		80-105km	90-100	80-105					
화자종속	G 화자	84.85	90.91	93.94				89/99	92.42 (183/198)
		95-110km	95-110	95-100					
	H 화자	90.91	93.94	93.94				94/99	
		80-95km	70-80	75-105					

2.5. 인식결과

오프라인 실험결과를 바탕으로 시스템을 설계, 구현한 후 실제 주행중인 차량에서 인식실험을 하였다. 그 결과 80km/h 이내의 환경에서는 화자독립 92.89%(화자 6인, 759단어), 화자종속 94.44%(6명화자, 990단어)의 인식율을 구하였고, 고속 주행환경에서는 【표 2】, 【표 3】과 같은 인식율을 구하였다.

Ⅲ. 결론

본 논문에서는 밴드패스 필터를 통해 잡음을 제거하고, 영교차율과 단구간 에너지를 이용하여 자동 음성구간 검출을 구현하였다. 기준모델은 DMS 모델을 사용하였고, 잡음환경에 강인성을 갖도록 하기 위하여 저속환경과 고속환경으로 나누었다. 인식 알고리즘은 OSDP[6]를 사용하여 실제 주행중인 차량환경에서 인식실험을 수행한 결과 시멘트 고속환경 화자독립 89.75%, 아스팔트 고속환경 화자독립 92.29%, 저속환경에서 92.89%의 인식율을 구하였다.

[참고문헌]

[1] L. R. Rabiner, M. R. Sambur, "An Algorithm for Determining the End Points of Isolated Utterances", The Bell System Technical Journal, Vol.54, No.2, pp297~315, February 1975

[2] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech" J. Acoust. Soc. Am. 87(4), pp1738~1752, April 1990

[3] 유봉근 외 "주행중인 자동차 환경에서의 음성인식 연구", 한국음향학회 학술발표대회 논문집 제17권 제1(s)호, pp47~50, July 1998

[4] 변용규, "DMS 모델을 이용한 단독어 인식에 관한 연구", 박사학위 논문, 광운대학교, 1990. 12

[5] 유봉근 외 "고속 주행중인 자동차 환경에서의 음성인식 연구", 제15회 음성통신 및 신호처리 워크샵 15권 제1호, pp65~69, August 1998

[6] Hermann Ney, "The Use of a One Stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No.2, pp263~271 April, 1984