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

실시간 윈도우 환경에서 DMS 모델을 이용한 자동 음성 제어 시스템에 관한 연구

A Study on the Automatic Speech Control System Using DMS model on Real-Time Windows Environment

남 동 선*, 김 순 협
(Dong-sun Nam*, Soon-hyop Kim)

요 약

본 논문은 인식 속도의 개선을 위해 단어의 지속시간에 따라 Section의 수를 변경한 가변 섹션 수 DMS모델을 사용한 실시간 인식 시스템을 연구하고 인식된 결과를 실제 수행하도록 하는 시스템을 구현하는 것이 목적이다. 이러한 윈도우 음성 제어 시스템 구현을 위해 음성의 자동 검출, 윈도우 제어 모듈 구현, 동적 모델 재구성을 이용하여 적용된 단어 단위 인식 시스템의 단점을 장점으로 수용하는 시스템을 구현하였고 본 시스템의 이름은 "VocManagerII"라 명명하였다. 구현된 시스템의 성능 평가 결과 인식 및 제어 수행 속도는 1초이내에 이루어지며 인식율은 66개의 기본 명령어에 대하여 화자 종속 99.36%, 화자 독립 99.08%의 좋은 인식율을 보여 주었다.

I. 서론

21세기 정보의 홍수 속에서 사람들은 다양한 정보 처리 서비스를 제공받을 수 있게 되었다. 이러한 서비스는 대부분 컴퓨터를 통하여 이루어지고 있다. 반면 일반 사람들은 컴퓨터를 사용하는데 대부분 미숙련자들이다. 이러한 사용자를 위해 새로운 인터페이스, 편리하고 쉬운 인터페이스의 연구가 많이 진행 중이다[1]. 그 중, 인간과 기계간의 가장 쉬운 의사 소통 도구로 인식되는 음성을 이용한 인터페이스에 관한 연구가 실용화를 위하여 전화망 서비스, 차량 제어 시스템, 윈도우 제어, 공장 자동화, 의료 기기 등의 분야에서 연구 중이다.

본 논문에서는 이러한 인간 기계 사이에 가장 편리하고 쉬운 인터페이스로 고려되어지는 음성을 이용하여 윈도우 시스템을 제어함으로 사용자에게 편의성, 작업의 효율성, 작업의 용이성 등의 장점을 제공하기 위해 윈도우 시스템에서 작동하는 윈도우 음성 제어 시스템을 구현 및 설계하였다.

본 연구에서 구현된 시스템은 고립단어인식 시스템의 단점을 장점으로 수용할 수 있도록 하였으며, 사용모델은 Dynamic Multi-Section을 사용하였으며, 인식 알고리즘은 인간이 인식을 단어 인식 시간 성도의 시간에 따라해 준다 OneStage DP 방법을 이용하여 단어 인식에 사용하였다. 또한 특징 벡터로는 LPC, Mel-Cep, PLP 각각 13차의 특징 벡터를 비교 실험하여 PLP[2]등을 인식 시스템의 특성 벡터로 사용하였으며, 실시간 음성의 끝구간 검출을 위해 300msec 마다 임계값을 재설정 하였고 실대 에너지와 영 교차율 값을 이용하였으며, 실제 윈도우에서 음성을 입력받을 때 발생 할 수 있는 주변 잡음의 처리를 위해 입력 신호에 대한 최대 Peak Amplitude를 사용하였다. 또한 사용자의 작업 공간을 최대한 침범하지 않

고 사용자에게 인식 시스템의 현재 상태 등의 정보 제공을 위해 최소 공간을 활용하는 인터페이스를 적용하여 시스템을 구현하였다.

Ⅱ. 본론

1. 자동 음성 구간 검출[2][6]

본 논문에서 구현된 자동 음성 구간 검출 알고리듬은 크게 기본 신호 처리 구간(Basic 1,2), 음성 시작 버퍼 검출 구간, 음성 끝 버퍼 검출 구간의 3 부분으로 나뉘어 진다. 기본 신호 처리 구간에서는 DC Offset을 제거하기 위해 입력 버퍼의 최초 5 Frame 값의 Mean Value를 계산하여 이후 데이터로부터 차감 해준다. 또한 임계값 설정을 위해 입력 신호로부터 절대 에너지와 ZCR을 계산하고 이전 입력 버퍼에서 설정된 에너지와 비교하여 음성의 구간검출에 사용된다. 음성 시작 구간 검출에 사용되는 변수는 PreEnergy(이전 버퍼의 평균 절대 에너지), Energy (현재 버퍼의 평균 절대 에너지), EnergyBig(현재 버퍼에서 설정된 임계값), ZCRSum(현재 버퍼에서 계산된 ZCR의 총 합), ZCRThreh(이전 버퍼의 ZCR로부터 계산된 임계값)이고 다음 조건은 사용된 음성 시작 구간 검출 조건을 나타낸다.

Condition 1.
(PreEnergy > Energy) ∩ (EnergyBig > 750)..(1)

Condition 2.
(PreEnergy > Energy) ∩
(ZCRSum < ZCRThreh)..(2)

위의 조건을 만족하여 음성의 시작 구간을 검출 후 끝점 검출은 간단하게 이전 구간의 절대 에너지로 설정된 임계값을 이용하여 예비 검출, 확인 검출의 두 부분으로 분리되어 처리된다. 또한 실제 윈도우를 사용자가 사용할 경우 실험 환경에서 발생되는 주변 잡음이 있기 마련인데, 이러한 잡음을 처리하기 위해 현재 입력되는 입력 신호의 평균 Peak Amplitude와 최대 Peak Amplitude를 이용하였다. 그림 1은 입력 되는 버퍼를 이용하여 음성의 시작 부분이 포함된 버퍼를 검출하는 처리 과정이며 그림 2는 입력이 시작된 후 계속적으로 입력되는 버퍼에서 음성 신

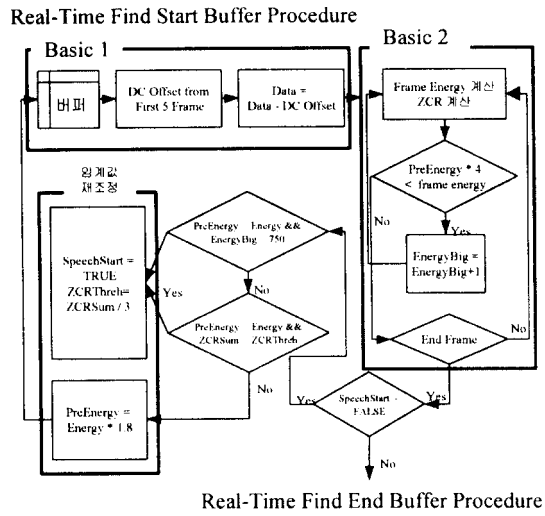

그림 1. 음성 시작 부분 검출 과정

호의 종료 부분을 찾아내는 알고리듬에 관한 처리 과정을 나타낸다. 구현된 처리 과정 수행 결과

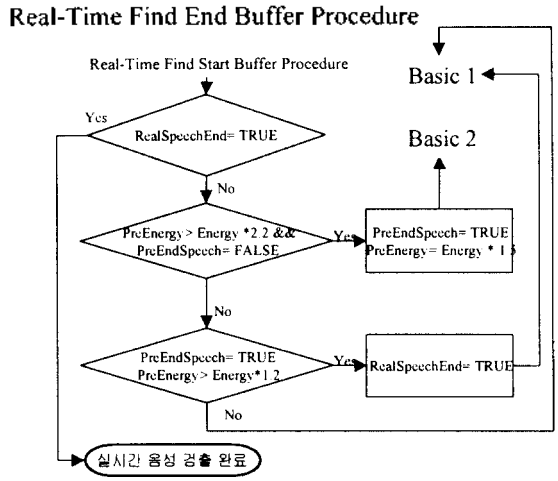

그림 2. 음성 종료 부분 검출 과정

음성 검출 시간은 발성 종료후 600msec 이내에 종료된다.

2. 가변 수 섹션 DMS 모델

인간은 Syllable 길이의 음성으로부터 정보를 얻어 이를 이해하고 의사 소통 정보로 이용한다.[5]

일반적으로 DMS 모델[3][4] 생성 시 섹션의 수를 모든 인식 대상 단어에 대하여 고정적으로 사용한다. 그러나 본 논문에서는 이러한 부분에서 발생되는 불필요한 계산 시간과 메모리를 줄이고,

궁극적으로는 인식 시간을 줄이기 위하여 인식 단어의 지속 시간에 따라 섹션의 수를 가변적으로 설정하는 가변 수 섹션 DMS 모델을 제안한다.

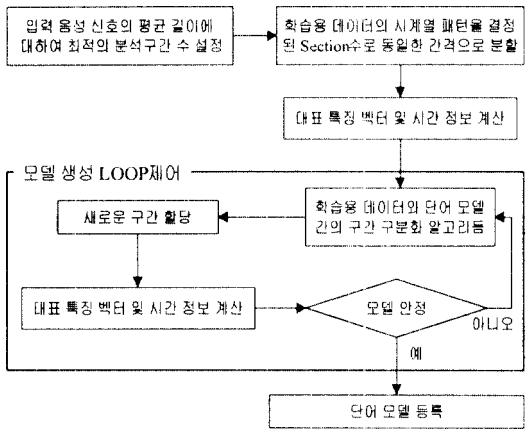

그림 4. 가변 섹션 모델 생성 과정

제안된 모델은 66개의 윈도우 제어 명령어의 DMS 모델로 구성되어 있으며 특징 벡터로 PLP 13차를 사용한다. 66개의 명령어는 400msec~1200msec의 지속 시간을 가진 단어들로 구성된다. 모델은 다중 화자가 발성한 단어를 이용하여 생성하고 각 단어에 대한 섹션 수 결정은 각 단어들의 지속 시간의 평균을 실험 결과로 나타난 한 섹션의 크기 45msec로 나누어 결정한다.

j 번째 단어의 섹션 수 : $\frac{\sum_{n=1}^{N} DT(n)_j}{T}$ .....(3)

DT : 입력 신호의 지속 시간

j : 발성 단어

n : 발성 화자

T : 할당된 섹션 당 지속시간 길이

다음 표1.은 구현된 음성 제어 시스템에서 인식 가능한 대상 단어의 다중 화자 발성 데이터에 대한 지속 시간과 그에 따라 위의 식 (3)에 의해 결정된 결과를 나타낸다.

표 1. 인식 대상 단어 당 섹션 수

음절 수	지속 시간	결정된 섹션 수
1, 2 음절	400~700msec	9 Section
3, 4 음절	750~1150msec	15 Section
5, 6 음절	1000~1200msec	20 Section

Ⅲ. 구현된 음성 제어 시스템

1. 사용자 인터페이스

구현된 시스템의 인터페이스는 사용자에게 편하도록 최소한의 사용자 작업 윈도우의 공간을

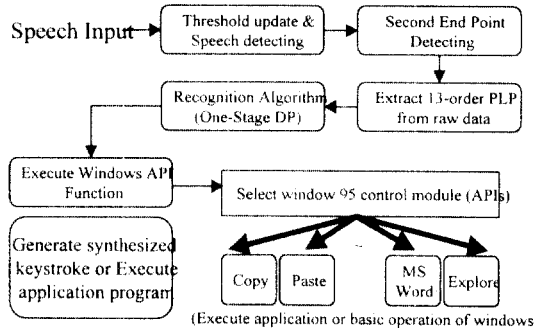

그림 5. 시스템 전체 구성도

차지하도록 구성하였으며 사용자에게 현재 인식 가능한 인식 대상 단어 및 모든 상황에서 인식 가능한 대상 단어들으로 분류하여 사용자에게 보여준다.

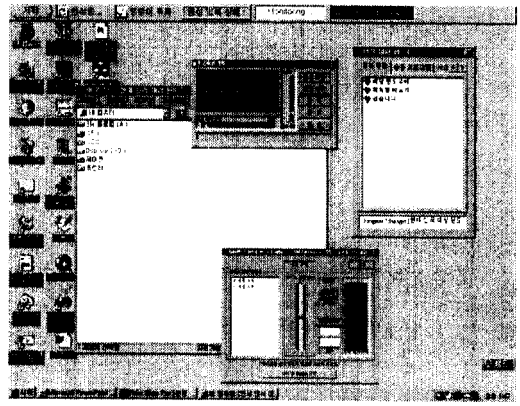

그림 6. 구현된 시스템 인터페이스

Ⅳ. 실험 및 고찰

1. DB 구축

실험에 사용한 음성 DB는 남성 화자 10인이 각각 3회 발성한 데이터를 이용하여 5인의 2회 발성 데이터는 모델 생성에 1회는 종속 화자 인식 실험을 위하여 사용하였으며, 나머지 5인의 데이터를 이용하여 화자 독립 인식 실험을 수행하

였다. 다음 표 2 는 음성 입력 설정 상황을 나타낸다.

표 2. 입력 데이터 설정

설정 내용	값
Sample 주파수	11.025 (KHz)
Channel 수	Mono (Channel 1)
양자화 bit 수	16 bits

2. 실험 결과

인식 실험은 첫째, 파라메타 비교 실험, 둘째, 섹션에 따른 실험, 셋째, 구현된 시스템에서의 실험으로 나누어 수행된다.

표 3. 화자 종속 파라메타 비교 실험 결과

	화자A	화자B	화자C	화자D	화자E	평 균
LPC15	92.42	93.93	96.96	96.96	93.93	94.84
LPC20	98.48	95.45	100.0	95.45	96.96	97.27
Mel15	96.96	93.93	98.48	95.45	95.45	96.06
Mel20	98.48	96.96	96.96	98.48	100.0	98.00
PLP15	100.0	100.0	100.0	96.96	98.48	99.09
PLP20	100.0	100.0	100.0	96.96	100.0	99.39
평 균	97.72	96.71	98.73	96.71	97.47	97.47

표 4. 섹션 수에 따른 인식율 비교 실험 (종속)

	5 섹 션	6 섹 션	7 섹 션	8 섹 션	9 섹 션	10 섹 션	11 섹 션	12 섹 션	13 섹 션	14 섹 션
A	27.2	57.5	66.6	75.7	92.4	89.3	95.4	96.9	100	100
B	37.8	50.0	62.1	75.7	86.3	89.3	93.9	96.9	100	100
C	34.8	48.4	68.1	80.3	89.3	93.9	96.9	100.	100	100
D	36.3	42.4	60.6	74.2	83.3	86.3	89.3	96.9	95.4	95.4
E	34.8	40.9	65.1	72.7	87.8	90.9	93.9	93.9	95.4	96.9
평 균	34.2	47.8	64.5	75.7	87.8	89.9	93.9	97.2	98.7	99.0

표 3.의 결과 LPC, Mel Cep, PLP 특징 벡터 각각 13차를 이용하여 DMS 모델의 섹션 수를 15, 20 섹션으로 변화시키면서 비교 실험을 수행하였다. 실험 결과 PLP - 13차 DMS 20 섹션의 경우 화자 종속으로 99.39의 인식율로 가장 우수한 결과를 나타내었다. 또한 표 4.의 결과를 이용하여 각 단어당 섹션의 수를 결정하는데 사용하였다.

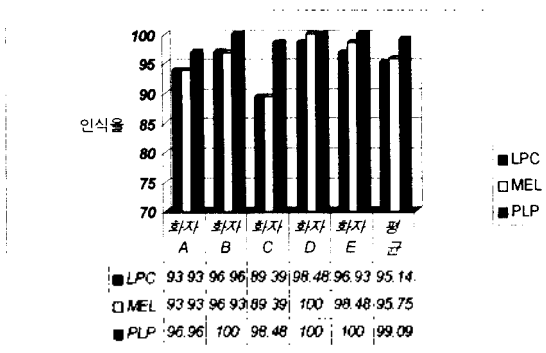

그림 7. 제안된 모델과 PLP 13차 인식 (독립)

V. 결과

본 연구에서는 기존 DMS 모델 생성에 고정식으로 사용하던 섹션의 수를 인식 대상 단어의 시속 시간에 따라 변경되는 가변 섹션을 적용하는 모델을 제안하였다. 본 시스템에서 사용한 인식 알고리듬과 파라메타를 이용하여 인식율은 거의 변화가 없었지만 인식 시간적 측면에서는 약 20% 개선되었다. 인식율은 66개의 제어 명령에 대하여 화자 독립 99.08%의 인식율을 얻었으며 윈도우에서 음성을 추가적인 인터페이스 수단으로 사용 가능하도록 음성 제어 시스템을 구현 하였다.

참고 문헌

[1] L.R. Rabiner, B.H. Juang, " Fundamentals of Speech Recognition" ,Prentice Hall, 1993.

[2] L.R. Rabiner, M.R. Sambur, " An Algorithm for Determining the Endpoints of Isolated Utterance", The Bell system Technical Journal, Vol. 54, No. 2, PP297-315, Feb. 1975.

[3] H. Sakoe, S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Transactions on communications, pp159-165, 1978.

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

[5] H. Hermansky, "Should Recognizers Have Ears?", Proc. ESCA Tutorial and Research Workshop on Robust Speech, 1994.

[6] 남동신, 이성숙, 이성권, 김순협, 이항섭, "음성 인식을 이용한 Windows 95 제어 시스템의 구현", 한국 음향 학회 학술발표회 논문집 제 17 권 1호 pp43-46, 1998. 6.