2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

CD-ROM Title 제어를 위한 음성인식 상용화 시스템에 관한 연구

이정숙*, 양진우**, 최갑근*, 김순협*
광운대학교 컴퓨터공학과*, 춘천기능대학 전자과**

A Study on Speech Recognition System Development for Control of CD-ROM Title

Jungsuk Lee*, Jinwoo Yang**, gabkeun Choi*, Soonhyob Kim*
Dept. of Electronics, Chunchon Polytechnic college**
Dept. of Computer Engineering Kwangwoon University*
e-mail: yjw@kopo.or.kr, kimsh@daisy.gwu.ac.kr

Abstract

본 논문은 상용화를 목적으로 하는 CD-ROM Title에 음성인식 시스템을 결합시킴으로써 컨텐츠의 검색과 제어를 쉽고 편리하게 하는데 목적이 있다. 콘텐서 마이크로 구축한 DB로 실험을 하였고, 또한 on-line 인터페이스와 병행하여 작업을 하였다. 본 논문은 상용화를 목적으로 하기 때문에 연령에 상관없이 인식이 가능해야 하므로 10대에서 50대에 걸쳐 데이터를 수집하였고, 인식 시에 전체 연령별로 인식률의 편차가 심하지 않게 하기 위해서 연령별로 데이터를 섞어서 생성하였다. 모델은 지속시간을 고려한 DMS 모델, 인식은 OSDP를 사용하였다.

I. 서 론

21세기는 지식 정보화 시대로 멀티미디어 컨텐츠 산업이 급 성장 할 것으로 전망된다. 따라서, 멀티미디어 컨텐츠의 매체로서 CD-ROM Title의 시장도 급 성장할 것으로 보인다. 이러한 대중화를 위해, 기존의 마우스나 키보드 보다 편리하게 사용할 수 있는 Interface인 음성인식 기술을 CD-ROM Title 제작기술과 연계한다면 컴퓨터 사용의 사전지식이 없는 어린이나 노인 어느 누구나 쉽게 사용가능 할 것이다.

본 논문은 학교 졸업앨범과 회사홍보용 CD-ROM Title에 음성인식을 적용하여 기존에 마우스로 클릭하여 실행하던 Content Title을 음성과 병행하여 사용할 수 있도록 하고자 한다. 이러한 시스템은 상용화를 목적으로 하기 때문에 남녀노소 누구든지 인식되어야 한다. 지속시간을 고려한 DMS model을 사용하였고 음성 특징 파라미터로는 인지 선형 예측(Perceptual Linear Prediction; PLP) 13차를 사용하였다. 인식 알고리듬은 OSDP(One Stage Dynamic Programming)방법을 단독어 인식에 적용하여 사용하였다. 회사 홍보용 CD-ROM Title에 사용되는 명령어 48단어를 선정하였다. 이 논문의 구성은 다음과 같다. 2장에서 지속 시간을 고려한 DMS 모델 생성방법 3장에서 DB구축 및 CD-ROM TITLE 인터페이스 4장에서 결과 및 고찰을 맺는다.

II. DMS 모델

2.1 지속시간을 고려한 DMS 모델

일반적인 DMS 모델의 생성과정은 두 단계로 수행되어진다. 첫째는 구간을 동적으로 분할하는 구간 구분화 작업이고, 둘째는 구분된 구간들에 대해 각 구간의 대표 특징 벡터와 지속 시간 정보를 구하는 단계이다. 기존의 DMS모델은 구분하는 섹션의 수를 고정하여 음성신호의 지속 시간에 관계없이 설정하여 비 효율성을 지닌다. 그러나, 본 논문에서는 수행 속도 및 인식 성능 향상을 위해 DMS모델의 한 섹션 당 일정한 지속시간을 가진 음성 신호 할당하여 구간을 나누는 개선된 DMS모델을 사용한다. 이는 음성을 인식하기 위해 지속시간에 대하여 일정한 분석 단위를 할당하는 것이다. 가변 섹션 수의 결정은 모델 생성을 위해 사용된 화자들의 음성 데이터를 이용하여 각각 음성 명령어들의 평

균지속시간과 섹션별 인식 실험을 통해 얻어진 결과를 이용하여 섹션의 수를 가변적으로 결정하였다. 다음 표 1은 결정된 섹션 분할 결과를 나타내고 그림 1은 지속 시간 정보를 이용한 가변 Section 모델 생성과정을 나타낸다.

표 1. 제안된 DMS 모델의 가변 섹션 수

단어의 음절 수	지속 시간 분포	결정된 Section 수
1,2 음절	350 msec ~ 700 msec	9 Section
3,4 음절	750 msec ~ 1150 msec	15 Section
5,6 음절	1000 msec ~ 1200 msec	20 Section

"글","예","영"…"십","공","게임","그림" 등은 9 section으로 3, 4 음절의 단어는 15 section, 5, 6 음절의 단어는 20 section으로 DMS모델을 구성하였다.

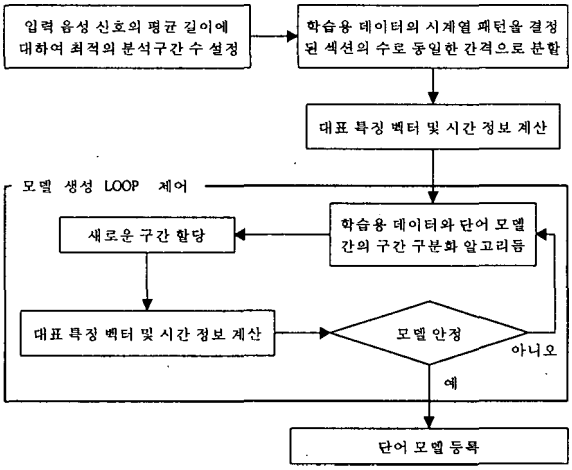

그림 1 지속 시간 정보를 이용한 가변 Section 모델 생성 과정

2.2 구간 구분화 알고리듬

DMS 모델 생성 시 최초로 입력된 학습용 데이터들은 시간 축 상에서 등분할 하여, 같은 구간으로 할당된 각 데이터의 프레임들의 특징 벡터들을 한데 모아 중심점을 계산하고 그 구간을 대표하는 특징 벡터를 구한다. 이 때 지속 시간 정보는 단어마다 그 수는 다르지만 간격은 구간마다 동일하게 분할했다는 것을 고려하여 각 구간의 마지막 프레임 수 들의 합을 학습용 데이터들의 전체 프레임수로 나누어서 구한다. 지속 시간 정보 $P(j)$는 다음과 같다.

$$P(j)=\frac{\sum_{m=1}^{M} e_m(j)}{\sum_{m=1}^{M} I_m} \quad ,1\le j\le J \quad \text{----------(1)}$$

위와 같이 구간의 정보를 구한 다음, 안정화되지 않은 각 초기 단어 모델과 동적 프로그래밍 매칭과정을 수행하고 백 트래킹 과정에 의해 구간의 경계선을 변경한다. 여기서 사용되는 동적 프로그래밍 알고리듬은 누적 거리 D에 지속시간 정보에 의한 거리 P를 포함시켜서 사용한다.

$$D(i,j)=d_v(t_i,m_j)+\min\begin{pmatrix} D(i-1,j), & (1<i\le I,\ 1<j\le J) \\ D(i-1,j-1)+P(j-1) & \end{pmatrix} \text{---(2)}$$

$$P(j)=W\times d_s(e(j),i) \quad \text{----------(3)}$$

$$d_s(e(j),i)=|p(j)\times I-i| \quad \text{----------(4)}$$

T : 학습용 데이터

i : 학습용 데이터의 프레임 번호 ($1\le i\le I$)

j : 단어 모델 M의 구간 Number

W : 지속 시간 정보의 차에 대한 가중치

(본 논문에서는 0.4를 사용)

M : 각 단어의 DMS 모델

J : 각 모델의 구간의 수

$e_m(j)$: 학습용 데이터의 j번째 구간의 마지막 프레임 Number

$d_v(t_i,m_j)$: 학습용 데이터의 i 번째 프레임의 특징 벡터 t_i와 모델 M의 j 번째 구간의 특징 벡터 m_j와의 Distance

$d_s(e(j),i)$: 모델의 j 번째 구간의 마지막 프레임의 Number와 학습용 데이터의 i 번째 프레임과의 차에 대한 절대값

III. DB 구축 및 실험

3.1 인식 대상 명령어

명령어는 회사 홍보용 CD-ROM Title 48단어로 아래 표 2에 나열하였다.

표 2 회사 홍보용 CD-ROM Title 명령어 목록

번호	명령어	번호	명령어	번호	명령어
1	상품안내	21	전화면	41	오
2	제품소개	22	종료	42	육
3	사용방법	23	예	43	칠
4	특징	24	아니오	44	팔
5	실행방법	25	프린트	45	구
6	문제해결	26	목차	46	십
7	게임	27	검색	47	공
8	오락	28	자동보기	48	다음
9	설치	29	끝내기	49	
10	직원소개	30	인터넷	50	
11	대표인사	31	메일	51	
12	회사소개	32	확대	52	
13	회사연혁	33	축소	53	
14	사업방향	34	메인	54	
15	협력회사	35	처음	55	
16	업무실적	36	영	56	
17	조직도	37	일	57	
18	연락처	38	이	58	
19	그림	39	삼	59	
20	글	40	사	60	

3.2 데이터 베이스 구축

아래 표 4의 조건으로 DB를 구축하였고 또한, 인터넷

표 4 입력 데이터의 설정

설 정 내 용	값
Sampling rate	11.025 KHz
Channel 수	Mono(Channel1)
양자화 bit 수	16 bit
마이크	콘덴서
환경	조용한 사무실 환경

각 DB는 남녀노소(10 ~ 50대)의 데이터를 54명이 3번 발성한 데이터를 수집하였고, 각 CD-ROM Title의 모델은 26명이 두 번 발성한 데이터로 모델을 구성하고 38명으로 인식 Test를 하였다.

표 5 전체 데이터 수집 현황 (명)

연령 / 성별	10대	20대	30대	40대	50대	합계
남	5	8	8	5	1	27
여	5	8	8	5	1	27

3.3 모델 생성

이 시스템은 실용화를 목적으로 함으로 주된 연령 사용자를 비롯한 대부분의 연령의 사용자가 모두 사용할 수 있어야 한다. 이 시스템은 컴퓨터의 사용방법에 대한 사전지식을 가지고 있지 않은 연령을 대상으로 한다. 따라서, 연령별로 모델을 만들어 나머지 연령의 인식실험을 해본 결과 즉, 예를 들면 20대 남성의 데이터로만 모델을 만들고 10대에서 50대까지의 연령데이터를 모두 인식 실험을 한 결과 30대 남성의 인식률은 비슷하게 나오지만 10대 남성이나 여성, 50대 남성, 여성의 경우는 인식률이 매우 저조하였다. 따라서, 주된 사용자만으로 모델을 만들면 연령별로 인식률의 차이 폭이 크므로 전체적인 연령 인식 데이터로 모델을 구성한다.

표 6 reference 구성 (명)

연령 / 성별	10대	20대	30대	40대	50대	합계
남	6	10	4	3	0	23
여	4	5	4	2	0	15

표 7 Test Pattern 구성 (명)

연령 / 성별	10대	20대	30대	40대	50대	합계
남	2	5	5	2	0	14
여	2	5	5	2	0	14

3.3 CD-ROM TITLE 인터페이스

본 논문에서 CD-ROM TITLE을 제어하기 위한 기술로 키보드를 제어하는 방법을 사용하였다. 먼저, CD-ROM TITLE은 키보드 제어가 가능하도록 제작해 놓고, 음성으로 인식된 결과에 따라 키보드 키를 자동으로 눌러 줌으로써 음성으로 실행이 가능하도록 한다.

기본적인 메시지 전달 창은 최상위 윈도우이다. 때문에 최상위 윈도우의 핸들을 획득하여야 한다. 최상위 윈도우 제어를 하기 위하여 GetForgroundWindow(), FindWindow()등의 함수를 사용하였다. 또한, 음성인식 시스템의 가장 중요하고 기본적인 기능은 음성 입출력 기능이라고 할 수 있다. 윈도우 운영체제에서는 이러한 멀티미디어 기능을 프로그래머가 쉽게 구현 할 수 있도록 MMSYSTEM이라는 라이브러리를 제공한다. MMSYSTEM이라는 라이브러리는 Low-level 함수, Mid-level 함수 그리고 High Level 함수로 나뉜다. 그러나 High-level 함수의 경우 사용자가 단지 음성 및 신호의 입력과 출력을 쉽게 할 수 있게 한 함수들로 음성 데이터를 이용하여 절대 에너지를 구하고 영교차율을 계산하고 특징 파라메터를 분석하는 데이터 조작이 가능하도록 함수를 제공하지는 않는다. 따라서, 데이터

에 대한 산술적 계산 등을 위해 Low-level 함수를 이용한 음성 입력과 출력 라이브러리를 구성하여 사용하였다.

IV. 실험 결과 및 고찰

4.1 CD-ROM Title 대상어휘 인식 흐름도

아래 그림 2는 대상어휘의 회사홍보용 CD-ROM Title·졸업 앨범용 회사홍보용 CD-ROM Title의 인식 흐름도를 보여준다.

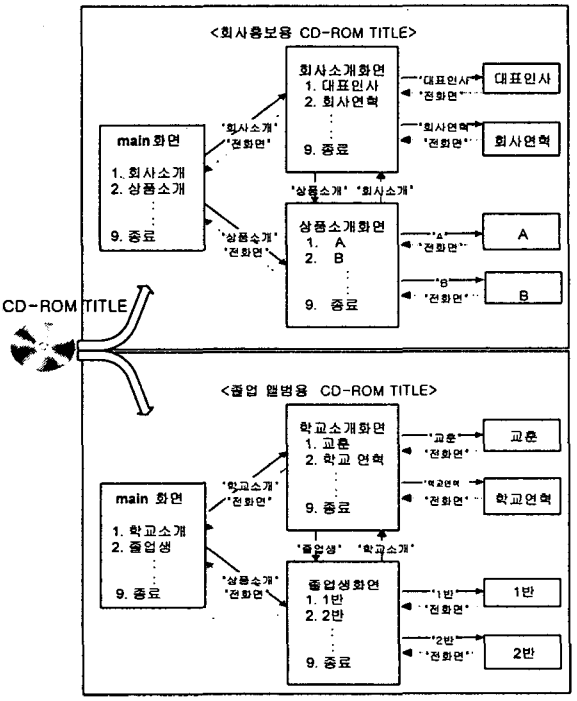

그림 2 CD-ROM Title 대상어휘 인식 흐름도

4.2 Off-line 실험결과

① 회사 홍보용

표 8 연령별 평균 인식률 (%)

연령 / 성별	10대	20대	30대	40대	평균
남	87.23	93.62	95.74	91.49	92.02
여	91.36	91.49	96.77	88.51	92.03

4.3 향후 계획

상용화 시스템을 위해서 마이크의 선택이 아주 중요하다. 본 실험에서는 콘덴서 마이크를 사용하여 실험을 하였는데, 인터넷 폰의 대중화로 일반 가정에서 널리 쓰는 마이크가 헤드 셋일 것이다. 따라서, 마이크를 헤드셋으로 데이터 베이스를 구축하여 위와 같은 실험을 다시 하고자 한다. 실험 대상(Test Pattern)을 100명으로 늘려서 실험 할 것이다. 또한, 인터페이스도 병행하여 구축하고자 한다. 인터페이스부분에서 고려할 점은 CD-ROM TITLE에서 음악과 설명이 나오는데, 이것과 동시에 음성인식 장치를 실행시켜서 사용할 때, 노트북에서는 아무런 문제없이 음성이 받아 들여 지던 것이 데스크 탑에서는 MMSYSTEM 장치가 충돌을 일으켰다. 이러한 작업을 보안하고, 또한 CD-ROM TITLE을 넣으면 자동으로 인식프로그램을 설치하고 CD-ROM TITLE을 띄워 줄 수 있도록 인터페이스 병행하여 시행하고자 한다.

V. 참고 문헌

[1] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals", Prentice-hall, 1978

[2] H. Hermansky, "Perceptual linear predictive(PLP) analysis of speech", J.Acoustical Society of America 87(4), pp1738-1752, Apirl 1990.

[3] 남동선, "윈도우 환경에서 DMS 모델을 이용한 음성제어 시스템에 관한 연구", 석사학위 논문, 광운대학교, 1998

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

[5] 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.