CD-ROM Title 컨텐츠의 검색과 제어를 위한 음성인식 시스템 개발에 관한 연구

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

A Study on Speech Recognition System Development for CD-ROM contents searching & manipulation

Jungsuk Lee*, Jinwoo Yang**, Jongmoon Woon*, 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 CD-ROM Title 컨텐츠의 검색과 제어를 위한 음성인식 시스템 개발을 목적으로 한다. 인식명령어는 학교 졸업앨범 또는 회사홍보용 CD-ROM Title 컨텐츠의 타이틀(상품안내, 회사소개, 업무실적 등)로 구성된다. 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 47단어와 학교 졸업앨범용 명령어(초·중·고 용- 58단어, 대학교 용 - 57단어)를 선정하였다. 이 논문의 구성은 다음과 같다. 2장에서 지속 시간을 고려한 DMS 모델 생성방법 3장에서 DB구축 및 실험과 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 평균지속시간과 섹션별 인식 실험을 통해 얻어진 결과를 이용하여 섹션의 수를 가변적으로 결정하였다. 다음 표 1은 결정된 섹션 분할 결과를 나타내고 그림 1은 지속시간 정보를 이용한 가변 Section 모델 생성과정을 나

타낸다.

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.

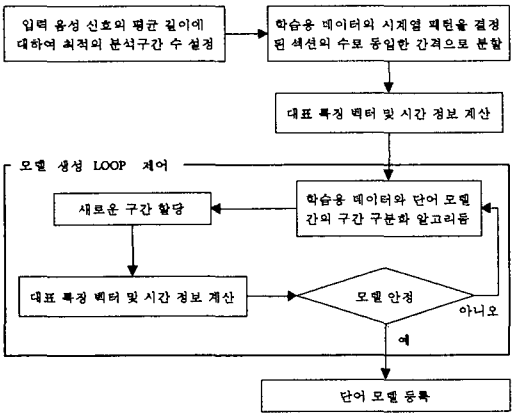

그림 1 지속 시간 정보를 이용한 가변 Section 모델 생성 과정

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.

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

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 프로그램밍 알고리듬은 누적 거리 D에 지속시간 정보에 의한 거리 P를 포함시켜서 사용한다.

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

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

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

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 크게 3종류로 나누고, 각각의 모델을 생성하여 실험한다.

첫째, 회사 홍보용 CD-ROM Title 명령어 목록으로 아래 표 2에 나열하였다.

둘째, 학교 졸업앨범 (대학교 용)은 표 3에 있다. 셋째, 학교 졸업앨범(초중고 용)은 아래 표 3에서 10, 11, 13번을 제외하고 "담임선생님", "소풍", "수학여행", "교장선생님", "선생님"이라는 단어를 추가한 것으로 구성되어 있다.

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	
9	설치	29	끝내기	49	
10	직원소개	30	인터넷	50	
11	대표인사	31	이 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	

표 3 학교 졸업 앨범 CD-ROM Title
명령어 목록(대학교)

번호	명령어	번호	명령어	번호	명령어
1	졸업생소개	21	교조	41	확대
2	단체사진	22	교목	42	축소
3	개인사진	23	교육목표	43	메인
4	낙서장	24	교가	44	처음
5	동영상	25	교훈탑	45	게임
6	추억	26	학교연혁	46	영
7	특별활동	27	학교연락처	47	일
8	축제	28	그림	48	이
9	주소록	29	글	49	삼
10	총장인사말	30	전화면	50	사
11	학장	31	종료	51	오
12	학과소개	32	예	52	육
13	교수소개	33	아니오	53	칠
14	학생소개	34	프린트	54	팔
15	동아리	35	목차	55	구
16	캠퍼스소개	36	검색	56	십
17	임직원소개	37	자동보기	57	공
18	졸업여행	38	끝내기	58	
19	교훈	39	인터넷	59	
20	교기	40	메일	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)

성별 \ 연령	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)

성별 \ 연령	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)

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

IV. 실험 결과 및 고찰

4.1 CD-ROM Title 대상어휘 인식 흐름도

아래 그림--은 대상어휘의 회사홍보용 ·졸업 앨범용 CD-ROM Title의 인식 흐름도를 보여준다.

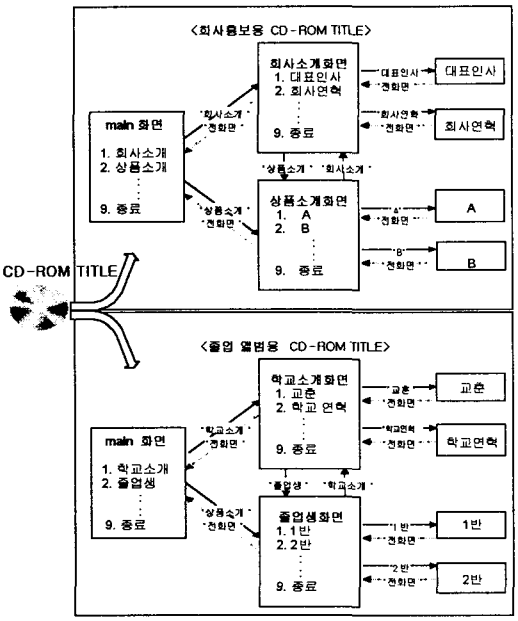

그림 2 CD-ROM Title 대상어휘 인식 흐름도

4.2 Off-line 실험결과

① 회사 홍보용

표 8 연령별 평균 인식률 (%)

성별 \ 연령	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

② 학교 졸업앨범(대학교 용)

표 9 연령별 평균 인식률 (%)

성별 \ 연령	10대	20대	30대	40대	평균
남	89.36	95.16	93.62	88.71	91.71
여	90.32	91.94	87.10	93.62	90.75

③ 학교 졸업앨범(초중고 용)

표 10 연령별 평균 인식률 (%)

성별 \ 연령	10대	20대	30대	40대	평균
남	95.75	93.62	91.49	89.36	92.5
여	93.55	93.62	90.32	87.23	91.18

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 구축하려고 한다. 데이터를 받기 위한 사전 테스트를 해본 결과, 기존의 콘덴서 마이크보다 잡음이 많고 아주 민감하였다. 또한, 헤드셋이라도 제품마다 소리의 입력 level, 잡음 정도 등의 다른 특성을 나타내었다. 따라서, 어떤 마이크를 사용하든지 적정 level내에 음성이 입력이 되게 하려면 먼저, 입력된 음성이 큰 것은 마이크 level을 자동으로 레벨을 낮춰줘서 입력의 크기를 줄여주고, 잡음은 BandpassFilter로 음성부분만을 제한하여 잡음을 제거하는 모듈이 필요할 것으로 보인다.

헤드셋으로 DB를 구축하고, 실험 대상(Test Pattern)을 100명으로 늘려서 실험 할 것이다.

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.