홍 기 형 교수지도
석사학위 청구논문

단순 특징 값과 촬영 각도에 따른 한국어 모음의 오디오 비주얼 인식에 관한 연구

2005

성신여자대학교 대학원
전 산 학 과
서 재 영

단순 특징 값과 촬영 각도에 따른 한국어 모음의 오디오 비주얼 인식에 관한 연구

홍 기 형 교수지도

이 논문을 석사학위논문으로 제출함

2004년 11월

성신여자대학교 대학원
전 산 학 과
서 재 영

인 준 서

서재영의 석사학위 논문으로 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논문개요

음성 인식은 잡음 환경에 매우 취약하며, 유사 발음의 단어에 대해서는 그 인식 신뢰도가 매우 저하된다. 음성 인식 결과와 발화시 입술의 움직임 정보를 이용한 입술 영상 인식 결과를 결합하여 음성 인식의 신뢰도를 높이기 위한 오디오-비주얼 인식이 시도 되고 있다. 본 논문에서는 다양한 각도의 영상으로부터 추출한 제한된 입술 특징 값을 사용한 오디오-비주얼 한국어 모음 인식 실험을 수행하여 촬영 각도에 따른 입술 특징 및 인식의 차이를 비교하였다. 각 모음 음성에 대해 발화자의 입술이 다문 상태와 비교하여 최대로 변할 때의 정지 영상으로부터 입술 특징을 추출 하였으며, 본 논문에서 추출한 입술 특징 값은 간단하며 효과적인 비용으로 얻어 질 수 있다. 이는 PDA나 스마트 폰과 같은 제한된 컴퓨팅 파워를 가지는 장치에서 매우 유용하다. 한국어 모음 음성 인식 오류를 입술 영상 인식 결과와 결합하기 위한 N-best rescoring 방법을 제시하고, 다양한 각도의 영상으로 추출한 입술 특징을 사용한 음성-입술정보의 결합 실험을 수행하여, 촬영 각도에 따른 입술 특징 및 인식의 차이를 비교하였다. 실험은 오디오 데이터 200개와 영상 데이터 1200개를 사용하여 HTK3.2[9]와 SNNS4.2[10]를 통해 이루어졌다. 실험 결과는 제안된 방법이 15° 이내의 각도 변화에서 오디오-비주얼 음성 인식에 효과적임을 보여준다.

목 차

그 림 목 차

표 목 차

Ⅰ. 서론

최근 간편한 이동성을 제공하고 소형화된 각종 장비가 빠른 속도로 보급되고 있다. 또한 모든 기기에 컴퓨팅 및 통신 기능을 집어넣어 언제, 어디서든 자유로운 커뮤니케이션이 가능하도록 하는 시도 역시 이루어지고 있다. 이러한 시스템을 유비쿼터스(Ubiquitous) 컴퓨팅이라 한다. 유비쿼터스 환경에서는 인간의 의사전달 수단이 비단 키보드와 마우스로만 한정되지 않는다. 특히 사람의 가장 자연스러운 의사 전달 수단은 음성으로 사람과 시스템 사이의 자연스러운 의사소통을 위해 음성 인터페이스의 사용 요구가 증가되고 있다.

음성 인터페이스의 실용화 작업에 있어 가장 큰 장애는 주변 환경의 노이즈로 인한 인식률 저하와 유사 단어 간의 인식률 저하 이다. 발화자의 입술 움직임을 포함하는 영상정보는 인식 정확도와 잡음에 강한 음성 인식 시스템에 매우 유용하다. 발화시 입술 움직임의 특징을 이용한 립리딩 인식을 음성 인식의 정확도를 위하여 함께 고려하는 연구는 오디오-비주얼 음성 인식(AVSR: Audio Visual Speech Recognition)이라고 한다[14][15][16].

그러나 과거 대부분의 AVSR 연구는 제한된 각도에서 수집된 영상 데이터와 복잡한 특징 값을 사용하여, 높은 컴퓨팅 파워를 요구하는 등 실용화가 어려운 환경에서 이루어져 온 경우가 많다. 특히, PDA나 스마트 폰과 같은 모바일 장비는 매우 제한된 컴퓨팅 파워를 가지고 있어 복잡한 영상 모델과 너무 많은 영상 특징 추출은 인식을 비현실적으로 느리게 한다. 또한 비주얼 인식 시 카메라가 항상 발화자의 정면에 위치한다는 보장이 없고 화자 역시 움직일 수 있기 때문에 정면의 입술 영상만을 고려하는 것은 비현실적이다.

따라서 본 연구에서는 낮은 컴퓨팅 환경에서도 실행 가능한 간단한 특징 값 추출 방법 연구 및 다양한 촬영 각도에 따른 인식 성능 검증 실험을 하였다. 실험을 위한 데이터베이스는 10명의 화자가 5가지의 한글 기본 모음을 4번씩 발화한 것으로 6가지(수평0° • 수직0°, 수평0° • 수직-15°, 수평0° • 수직+15°, 수평10° • 수직0°, 수평20° • 수직0°, 수평30° • 수직0°) 다양한 각도에서의 촬영을 통해 수집하였다.

수집한 데이터베이스를 활용하여 우선 음성 인식과 립리딩 각각에 대한 성능을 평가하였다. HTK 3.2 [9]를 사용하여 음성인식 실험을 수행하였으며 결과로는 N-best 신뢰도인 정규화된 LLR(Log Likelihood Ratio)를 사용하였다. 영상 부분에서는 SNNS 4.2 [10]을 사용하였으며, 입술 특징 추출 값을 입력으로 하는 Back propagation 신경망을 설계하였다.

음성만을 사용한 인식 결과의 정확도를 높이기 위하여 입술 영상의 립리딩 결과와 통합하는 방법을 제안하고, 촬영 각도 별로 통합 방법에 따른 인식 성능을 평가하고 분석하였다.

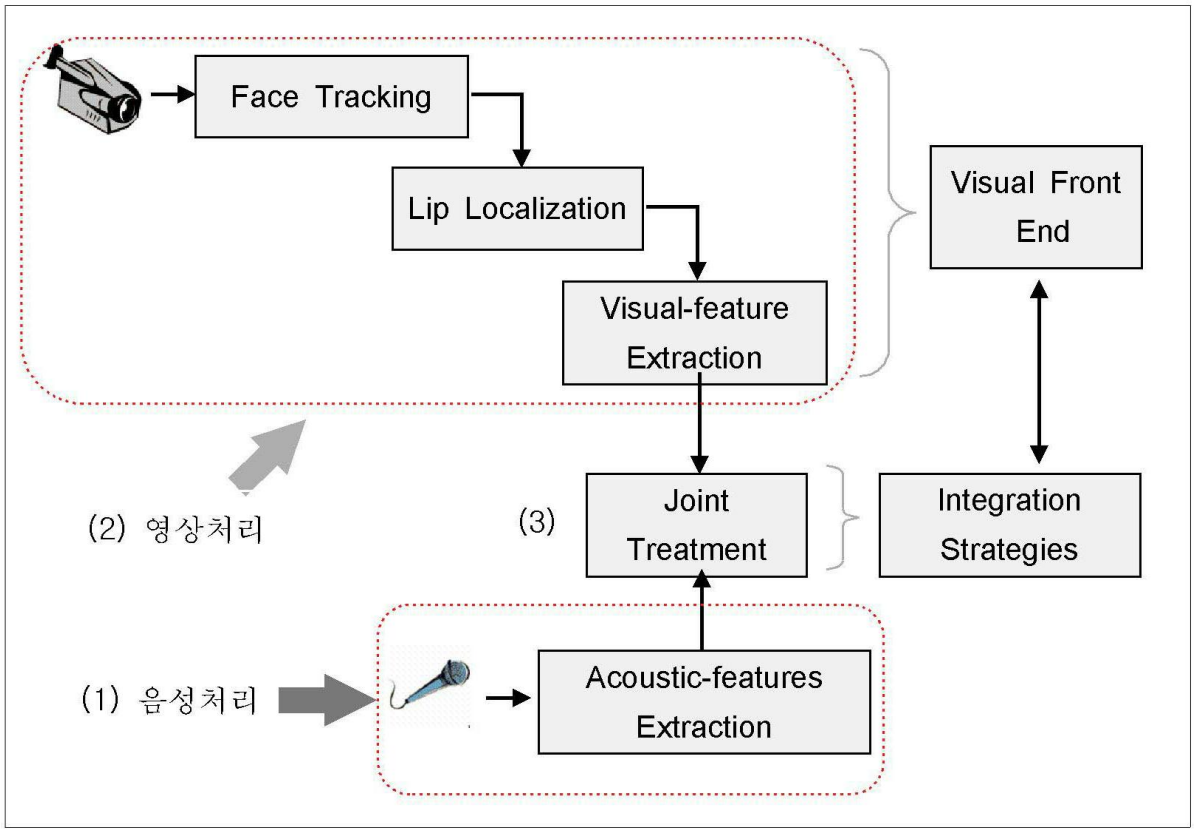
본 논문의 구성은 다음과 같다. Ⅱ장에서는 AVSR(Audio Visual Speech Recognition)의 개요 및 기존 연구 사례들에 대해 기술한다. Ⅲ장에서는 실험에 사용할 데이터베이스 수집환경에 대해 기술한다. Ⅳ장에서는 음성 인식 실험에 대해 Ⅴ장에서는 립리딩 실험에 대해 각각 기술 하였다. Ⅵ장에서는 Ⅳ장과 Ⅴ장에서 실험한 음성 인식과 립리딩 실험 결과를 통합하는 과정을 기술한다. 마지막으로 Ⅶ장에서 결론 및 향후 연구 과제를 제시하였다.

Ⅱ. 관련연구

1. 오디오-비주얼 음성 인식

오디오-비주얼 음성 인식(AVSR: Audio Visual Speech Recognition)은 음성 인식 수행 시 음성 신호와 발화자의 입술 모양 변화를 함께 고려하여 인식의 정확도를 향상시키는데 그 목적이 있다.

일반적인 AVSR의 과정은 [그림 1]과 같다. AVSR은 (1)음성처리와 (2)영상처리를 한 후 (3)통합하는 과정으로 구성된다. 영상 처리의 경우 우선 얼굴 위치를 추적 한 후, 입술 부위로 범위를 국한해서 영상 특징을 추출하는 여러 단계의 과정이 필요하다. 음성 처리의 경우는 음향학적 특징을 추출하는 단계를 거치게 된다. 마지막으로 음성과 영상 각각의 특징들을 결합하여 통합하는 과정을 거치게 된다.



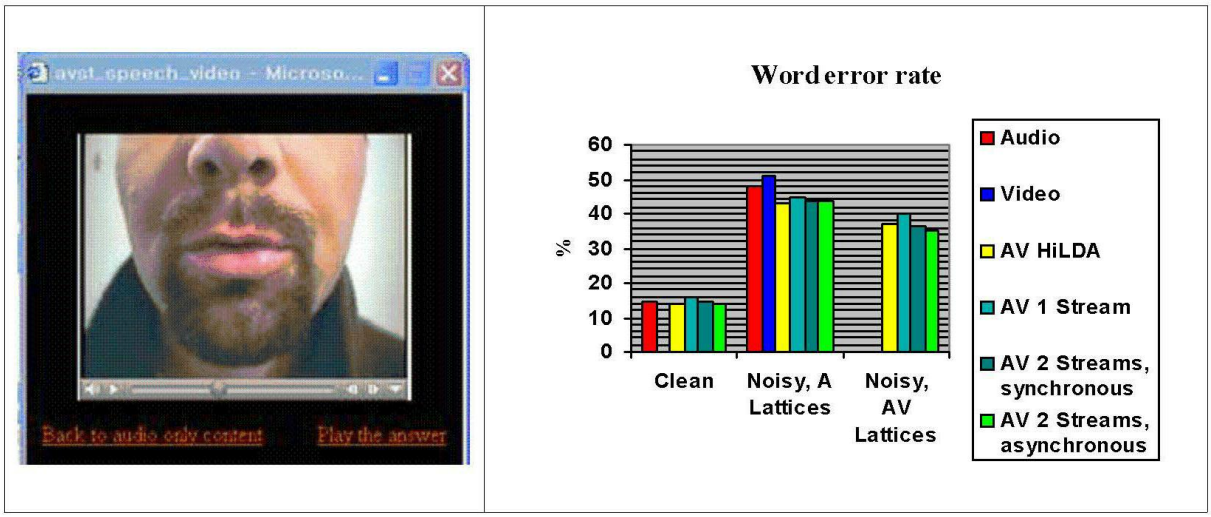
[그림 1] Audio-Visual speech system

2. 연구사례분석

본 절에서는 수집 방법과 수집 목적이 각기 다른 기존 오디오-비주얼 음성 인식 연구 사례에 대해 기술하고 각 사례를 분석하였다.

2.1 IBM [1]

IBM은 오디오-비주얼 음성 인식 실험을 위해 290명의 화자가 연속음 10,500단어를 50시간 동안 발화한 멀티모달 음성 DB를 수집하였다. IBM 오디오-비주얼 데이터베이스는 백그라운드 소음으로 약간의 컴퓨터 소리가 들어가 있다. 수집한 데이터베이스를 사용하여 인식 실험 수행 결과는 [그림 2]와 같다.



[그림 2] IBM 데모 시스템과 인식 실험 결과

[그림 2]의 왼쪽은 IBM 데모 시스템이고 오른쪽은 인식 실험 결과이다. 오디오만을 사용해서 인식한 결과 잡음이 없는 환경에서는 오인식율(WER: Word Error Rate)이 14%이고, 잡음이 있는 환경에서는 WER이 48.1%이다. 오디오만을 사용했을 때보다 오디오와 비디오를 함께 사용해서 인식을 한 경우, 10dB SNR(signal to noise rate)에서 WER이 27% 감소하였다.

2.2 CMU [2]

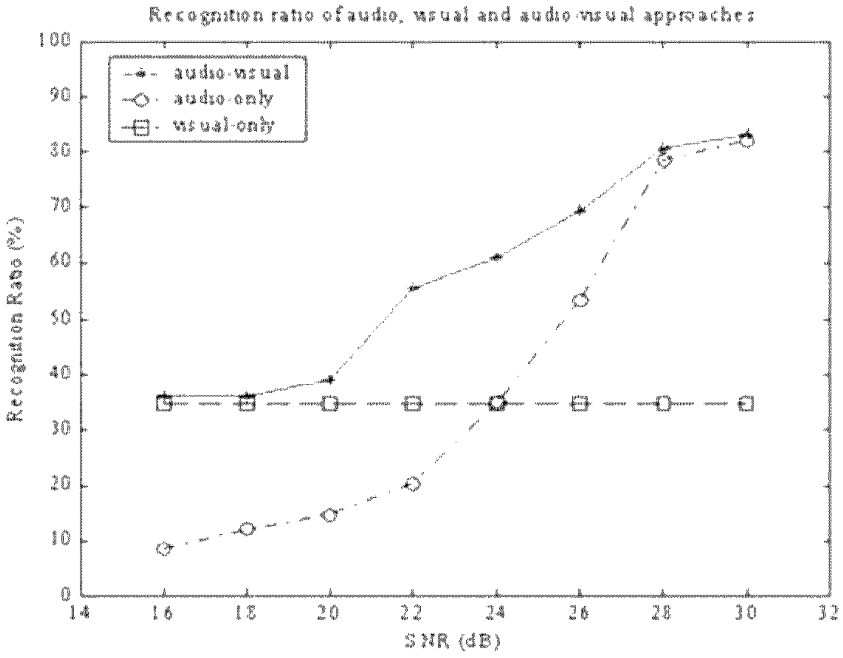
오디오-비주얼 음성 인식 실험을 위해 10 명의 발화자(남자 7명, 여자 3명)가 78단어를 10번씩 반복 발화하여 데이터베이스를 구축하였다. 발화 단어 리스트는 일상생활에서 흔히 사용하는 단어로 다음 [표 1]과 같다.

그룹	단어
For date/time	one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, thirty, forth, fifty, sixty, seventy, eighty, ninety, hundred, thousand, million, billion
For month	January, February, March, April, May, June, July, August, September, October, November, December
For day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Some additional word	morning, noon, afternoon, night, midnight, evening, AM, PM, now, next, last, yesterday, today, tomorrow, ago, after, before, from, for, through, until, till, that, this, day, month, week, year

[표 1] CMU-AMP Lab 단어 목록

녹화 환경은 soundproof 스튜디오에서 조명을 조절하고 블루 스크린을 배경으로 하여 촬영하였으며, 촬영 장비는 소니 디지털 캠코더와 tie-clip 마이크를 사용하여 DV 테이프에 녹화하였다. 소프트웨어로는 Radius MotoDV program과, Premiere를 사용하였다. 결과물은 Quicktime 파일의 두 가지 형태로, 해상도 720 × 480의 얼굴 전체 파일과, 해상도 216 × 264의 입술 부분만 따로 잘라내서 편집한 파일이다.

[그림 3]은 인식 실험 결과이다. 영상만을 사용한 인식결과는 소음에 관계없이 항상 일정한 인식률을 보인다. 그러나 잡음이 있는 환경에서는 (Signal to Noise Ratio가 높을수록 잡음이 적은 환경) 음성 정보만을 사용한 경우보다 음성과 영상 정보를 함께 사용한 경우에 인식률이 개선됐음을 볼 수 있다.

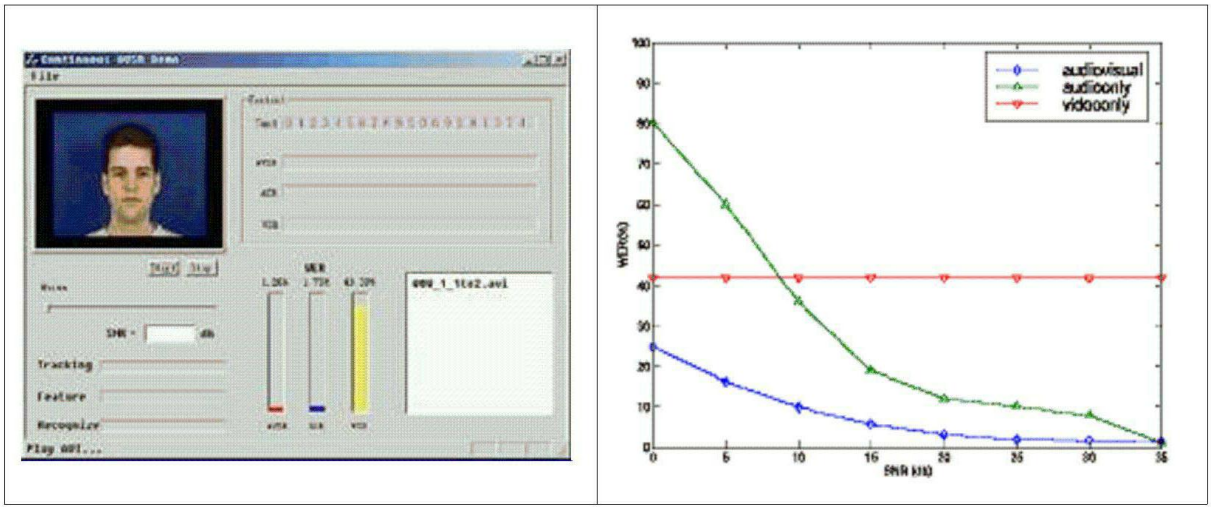


[그림 3] CMU의 데이터 인식 테스트 결과

2.3 Intel [3]

Intel에서는 오디오-비주얼 음성 인식 실험을 위해 CMU 데이터베이스를 그대로 가져다 사용했다. 시스템 개발 시 10번 반복 발화한 것 중 9개는 인식기 훈련에 사용하고 나머지 1개는 인식기 시험에 사용하였다.

[그림 4]에서 왼쪽은 Intel에서 개발한 시스템, 오른쪽은 인식 시험 결과를 각각 나타낸다. 인식 시험 결과는 0db의 SNR에서 음성만을 사용해서 인식했을 때 보다 음성과 영상 정보를 함께 사용하여 인식했을 경우, 55%의 WER 개선률을 보인다.



[그림 4] Intel 개발 시스템과 인식 테스트 결과

2.4 전남대 [4][5]

국내에서도 시청각 음성 합성과 오디오-비주얼 음성 인식을 목표로 하는 연구가 진행되었다. 두 경우 모두 마커를 부착해 입술 특징 정보를 더 쉽게 얻는 DB수집 방법을 시도하였다.

먼저 입술정보를 이용한 음성 특징 값 추정 및 음성인식 성능 향상을 목표로 하여 한국어 단모음 5개에 대한 데이터를 수집하였다 [4]. 20대 1인이 마커를 부착한 상태로 각 모음을 30회씩 발화하였다. 입 모양의 깊이 방향(z-축 방향) 정보를 얻기 위하여 거울을 이용하여 카메라의 영상에 얼굴의 정면 모습뿐만 아니라 옆모습 역시 투영되도록 하였다. 마커의 위치는 Tracker 툴을 사용하여 입술의 폭과 높이, 윗입술에서 턱까지의 거리, 코로부터 입술 위까지의 z 방향 거리, 코로부터 턱까지의 z 방향거리 등 8가지를 수집하였다.

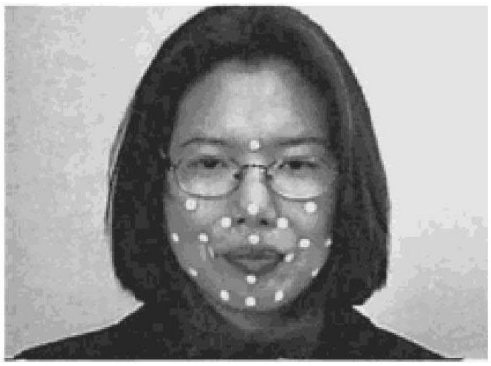
녹음 단어	한국어 단모음 5개 (아, 이, 우, 에, 오)
녹음화자 수 및 횟수	20대 1인, 각 30회 발성
A/D 변환	음성: 8kHz, 16bits 영상: 30frames/sec
마커	8개 사용 (입술의 폭과 높이, 윗입술에서 턱까지의 거리, 코로부터 입술 위까지의 z 방향 거리, 코로부터 턱까지의 z 방향거리)

[표 2] 마커를 이용한 음성 특징 파라미터 추정 및 성능 실험 환경

두 번째 경우는 오디오-비주얼 코퍼스에 기반한 Lip-Sync 알고리즘 개발을 위해 데이터베이스를 수집하였다. [4]에서와 같이 마커를 사용하였으며 데이터베이스 수집환경은 [표 3]과 같고 마커 부착 예는 [그림 5]와 같다.

녹음 단어	CVC syllable unit
Recorder	SONY 캠코더
A/D 변환	음성: 8kHz, 16bits 영상: 30frames/sec
마커	19개 4 markers around the lip contour + 7 markers along the jaw's contour + auto-reflective round markers

[표 3] 마커를 이용한 Lip-Sync 알고리즘 개발을 위한 DB 수집



[그림 5] 마커를 이용한 DB 수집 예

2.5 ATR [6]

ATR에서는 AVSR을 목적으로 한 것은 아니지만 얼굴 인식을 위해 다양한 각도를 고려하여 얼굴 인식 성능을 시험 하였다. 시험 결과는 [표 4]와 같다.

	Frontal face	10 deg. face	20 deg. face	45 deg. face
Face detection rate	98.90%	27.00%	11.30%	0.2%

[표 4] face-to-talk의 각도별 얼굴 검출 성능

얼굴 각도를 고려하여 얼굴 검출 성능의 시험 결과는, 정면의 경우 거의 완벽에 가까운 인식률을 보이나 얼굴 방향이 조금만 틀어져도 인식률의 현저한 저하를 보인다. 특히 각도가 45°만 되도 거의 얼굴 검출이 되지 않는다는 연구 결과가 나왔다.

2.6 분석

[표 5]는 위의 대표 사례들과 기타 사례들은 활용 목적 및 수집 방법에 따라 정리해 놓은 것이다.

	IBM	CMU	Intel	전남대 (IEICE'03)	전남대 (말소리'02)	AT&T	ATR
활용목적	AVSR	AVSR	AVSR	시청각 음성 합성	AVSR	Face recognition	Face detection
화자 수	290	10	10	1	1	40	68
단어셋	10500 단어	78단어	78단어	33개의 음절	한국어 단모음 5개	-	-
마커사용	-	-	-	19개	8개	-	-
앵글변화	-	-	-	-	-	10 poses	13 poses

[표 5] 국내·외 사례 분석

각 사례들을 분석 해본 결과 실용화가 어려운 환경에서 몇몇 연구가 이루어 졌음을 알 수 있다. 예를 들면 제한된 각도(정면)에서만 DB 수집이 이루어졌거나, 마커를 부착했거나, 복잡한 특징 값 및 높은 컴퓨팅 파워를 요구하는 특징들을 사용하였다. 그러나 실용화를 위해서는 낮은 컴퓨팅 환경에서도 실행 가능한 간단한 특징 값 추출 방법과 촬영 각도에 따른 인식 성능 검증이 필요하다.

본 논문에서는 멀티모달 음성 DB를 실제 응용 분야에 활용한다는 전제하에 연구가 이루어진 것이기 때문에 비실용적인 마커들은 사용하지 않는다. 또한 [표 4]에서의 결과가 나타내듯이 정면에서 45°이상 틀어진 영상에서의 얼굴 검출이 현재 기술 수준으로는 인식률이 매우 낮으므로, 본 논문에서는 입술 인식률 실험을 수행하기위해 30°이하의 각도만을 고려하여 DB 수집과 인식 실험을 진행 하였다.

Ⅲ. 데이터베이스 수집

1. 수집 환경 및 시스템 구성

다양한 각도에 따른 오디오 비주얼 음성 인식 실험을 위해 자체적으로 데이터베이스를 수집하였다. 다음 사항들을 고려하여 최종적인 DB 수집 환경과 시스템을 구성하였다.

첫째, 다수의 카메라를 사용한 여러 각도의 동영상 동시 촬영 방법의 다양한 시도를 통한 효율인 카메라의 수 및 촬영 각도를 결정한다.

둘째, 효과적인 카메라 해상도(raw data) 및 동영상•음성 코딩 스펙과 디지털화 방안을 고려한다.

셋째, 효과적인 동영상•음성 녹음의 동기화 방법 및 DB 수집 프로세스를 고려한다.

1.1 수집 환경

방음 시설이 잘 된 음향 스튜디오 부스 내에서 수집 작업을 하였으며, 사용한 장비 사양은 [표 6]과 같다.

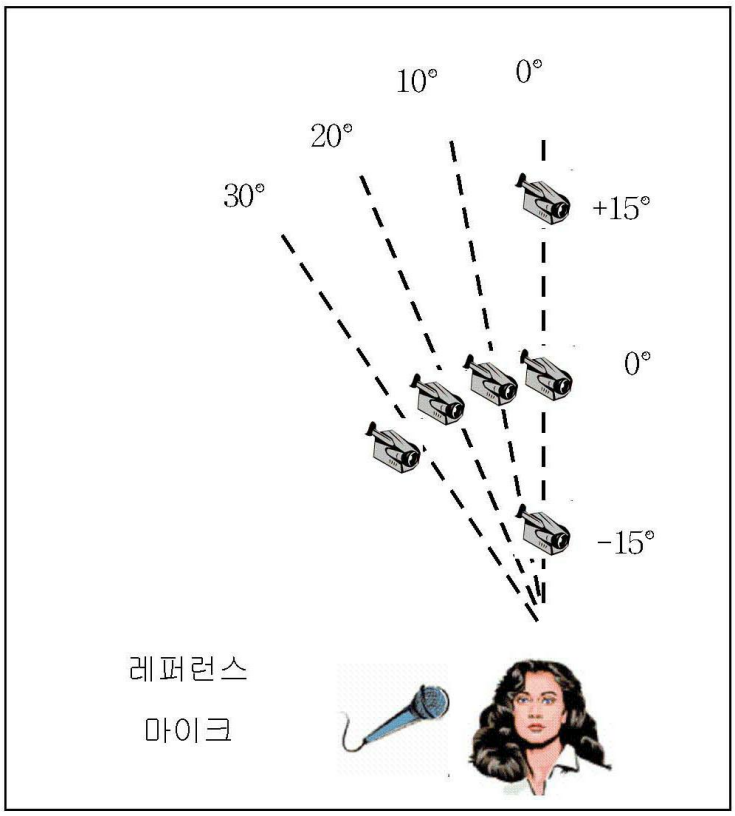
장소	음향 스튜디오 부스 (230cm x 230cm)
비디오	Sony DCR-VX 2000 x 6대, Sony MiniDV Tape
오디오	캠코더 내장용 마이크, SHURE-SM57
조명	음향 스튜디오 기본 조명, 임시조명
배경	블루 스크린

[표 6] 데이터베이스 수집 장비 사양

영상 데이터 수집을 위해 Sony DCR-VX 2000을 6대 사용하여, Sony MiniDV에 녹화하였다. 캠코더 여섯 대는 수평0° • 수직+15°, 수평0° • 수직0°, 수평0° • 수직-15°, 수평10° • 수직0°, 수평20° • 수직0°, 수평30° • 수직0° 형태로 배치 시켰다.

음성 데이터는 캠코더에 내장된 마이크와 SHURE-SM57 사용하여 방향성이 고려된 데이터와 레퍼런스 데이터 두 가지 종류로 수집하였다.

조명은 음향 스튜디오에 설치되어있는 기본 조명을 사용하고 추가로 임시 조명을 설치하여 화자의 얼굴에 최대한 그림자가 지지 않도록 하였다. 임시 조명은 정면(200W × 2개)에 1 세트, 좌 • 우(100W × 2개)에 2 세트, 하단(60W × 1개)에 1 세트, 총 4 세트를 설치 시켰다. 다음 [그림 6]은 녹음 시스템 구성도이다.



[그림 6] 데이터 수집 시스템 구성도

1.2 단어 선정

발화할 단어로는 한국어 기본 모음 다섯 가지인 '아', '이', '우', '에', '오' 다섯 가지를 선정하였다.

1.3 화자 선정

표준어 수집을 위해 화자의 고등학교 이후 거주지를 서울/경기로 제한한 20•30대 남•여 10명의 화자를 선정했다.

성별	남성	여성	합계
비율	40%	60%	100%

[표 7] 성별 화자 비율

연령대별	20대	30대	합계
비율	90%	10%	100%

[표 8] 연령대별 화자 비율

화자 정보는 화자 일련번호, 이니셜 3자리, 성별, 나이, 출생지, 성장 지역의 기준이 되는 12세 이전 성장지, 현 거주지(시, 군 단위까지), 현 거주지 거주 기간(년 단위), 직업 및 업태에 대한 정보(학생/직장인/주부/etc.), 부/모 출생지로 구성된다. 다음 [표 9]는 화자 정보 테이블이다.

번호	이니셜	성별	나이	출생지	12세 이전 성장지	현 거주지		직업	출생지	
						지역	기간		부	모
1	lsj	여	31	서울	서울	서울	31	직장인	충남	충남
2	smj	여	22	서울	경기	서울	11	학생	경기	서울
3	nek	여	23	서울	서울	서울	20	학생	전라도	전라도
4	cnc	남	25	서울	서울	서울	23	학생	전북	전북
5	lcs	남	25	서울	서울	서울	23	학생	전라도	전라도
6	ljr	남	25	경기	경기	서울	6	학생	서울	전라도
7	ljy	여	23	서울	서울	서울	23	학생	서울	서울
8	pyy	여	22	서울	충남	경기	10	학생	서울	서울
9	lmj	여	22	서울	서울	경기	15	학생	서울	충청도
10	kjh	남	27	서울	서울	서울	27	학생	전라도	전라도

[표 9] 화자정보

2. 데이터 가공

2.1 데이터 캡쳐

구성된 시스템에서 촬영을 통해 MiniDV Tape에 녹화된 데이터는 캡쳐 과정을 거쳐 6개의 avi 파일과 6개의 wav파일로 가공된다. MiniDV Tape을 DCR-VX 2000에서 IEEE 1394 포트를 통해 PC와 연결하여 직접 캡쳐한다.

데이터를 캡쳐하기 전에 적절한 설정 값을 정하기 위해 Keyframe에 따른 용량 변화를 테스트 해보았다. 테스트 결과 캡쳐 시 MPEG-4 V1으로 파일을 압축했을 때 Keyframe 설정 값에 의해 파일 용량이 크게 영향을 받지 않았다([표 10] 참고).

발화자		khj
원본(Audio: 48000Hz, 16bit, stereo Image: 720 x 480, 29.97fps, uncompressed)		Time 01:01
		File size 220M
압축형식: MPEG-4 V1 Image: 720 * 480, 30fps Audio: 48000Hz, 16bit, Stereo	Keyframe every 8 sec	34.4M
	Keyframe every 7 sec	34.4M
	Keyframe every 6 sec	34.4M
	Keyframe every 5 sec	34.4M
	Keyframe every 4 sec	34.4M
	Keyframe every 3 sec	34.4M
	Keyframe every 2 sec	34.4M
	Keyframe every 1 sec	34.4M

[표 10] Keyframe 설정에 따른 용량 변화

따라서 양질의 영상 DB를 위해 이미지의 Keyframe 값을 최대로 설정하였다. 압축은 MPEG-4 V1을 적용하고 frame 값은 30frame에 해상도는 720 x 480으로 설정하였다.

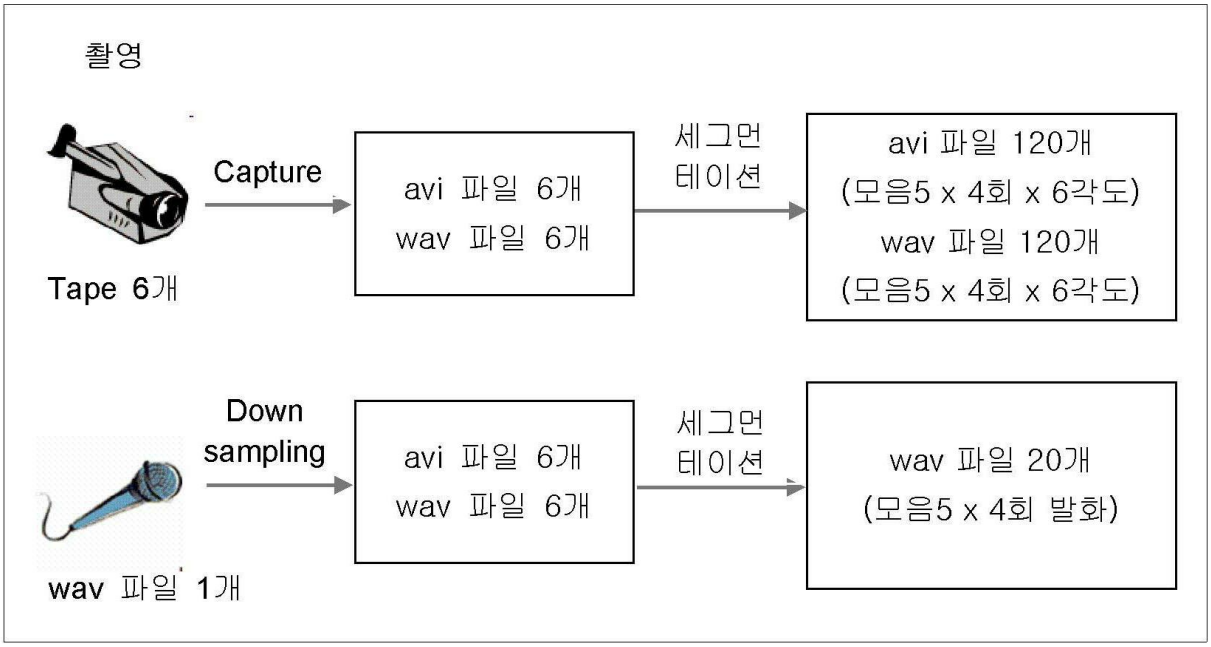
오디오 파일은 48 KHz, 16bit, mono상태에서 MiniDV Tape에서 각 카메라 별로 방향성이 고려된 6개의 wav파일과 외장형 마이크와 SoundForge를 사용해서 1개의 wav파일 총 7개의 wav파일을 가공하였다. 그러나 Premier에서는 PCM방식을 지원하지 않기 때문에 SoundForge에서 별도의 DownSampling 작업이 필요하였다.

비디오	소프트웨어	Premier 6.5
	압축	MPEG-4, key-frame 최대값
	설정	해상도 720 x 480, 30 frame
오디오	소프트웨어	SoundForge
	설정	48 KHz, 16bit, mono

[표 11] 데이터 캡쳐 설정

2.2 세그먼테이션

가공된 6개의 AVI파일과 7개의 wav파일들로부터 각 모음들을 세그먼테이션 하는 과정이 필요하다. 실험에 사용할 레퍼런스 wav 파일 외에도 각 카메라 별로 방향성이 고려된 6개의 wav 파일은 추후 사용을 위해 함께 세그먼테이션 했다. 세그먼테이션 작업은 Premier 6.5에서 이루어졌으며 각 모음의 음성 신호를 기준으로 해서 앞뒤 각각 0.2초의 silence 구간을 두고 잘라내어 다시 exporting하였다.

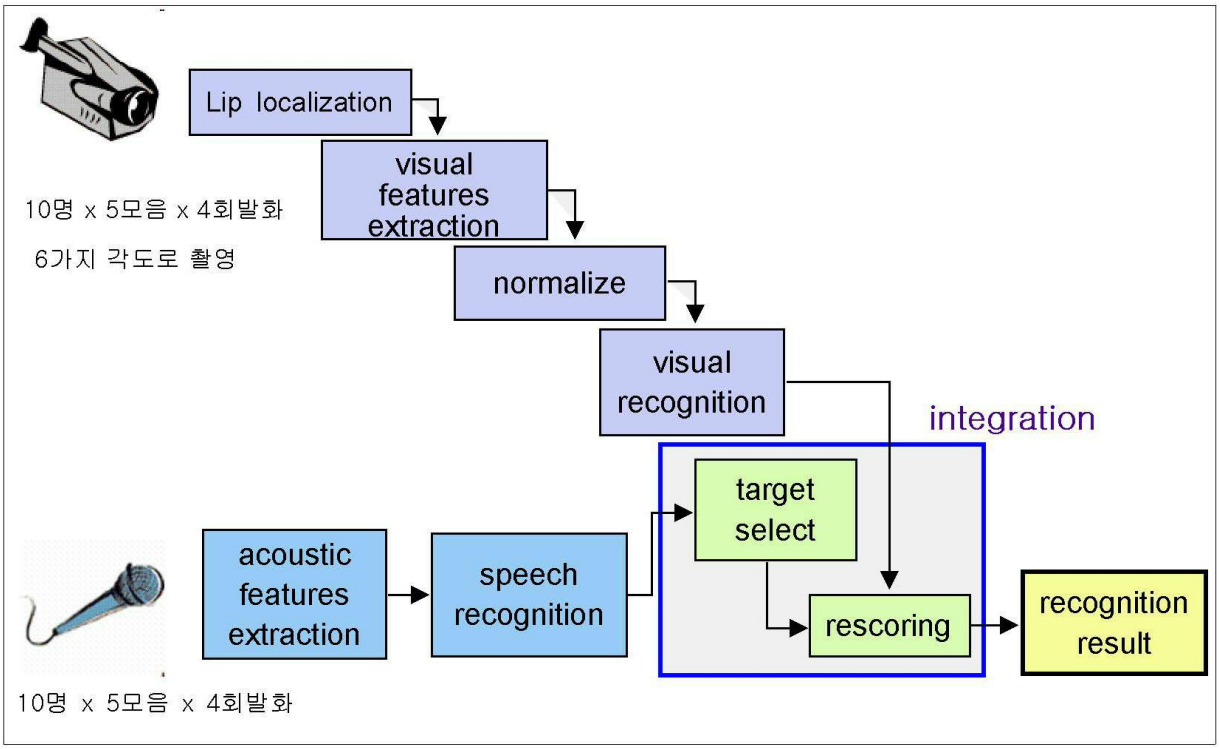


[그림 7] 데이터 가공 과정

모음 5개를 4번씩 10명이 발화한 6가지 각도에 따른 avi 파일 1200(5 × 4 × 10 × 6)개와 wav 파일 1200개로 세그먼테이션 하였다. 또한 별도의 레퍼런스 wav 파일 역시 총 200(5모음 × 4번 발화 × 10명)개의 파일로 세그먼테이션 하였다. 데이터 가공 결과 총 파일 개수 2600개이다.

Ⅳ. 실험 시스템 구성

본 논문에서는 다음 [그림 8]과 같은 오디오 비주얼 인식 시스템을 구성하였다.



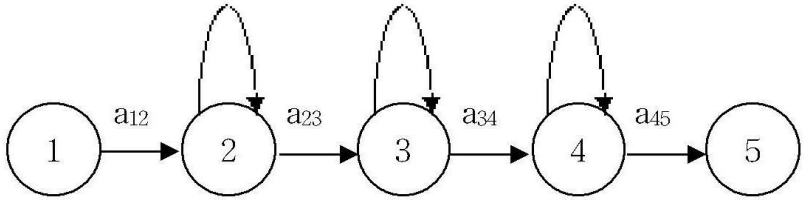
[그림 8] 실험 시스템

먼저 촬영 데이터에서 이미 입술 부분만을 따로 가공한 이미지 파일 1200개(10명 × 5모음 × 4회 발화 × 6각도)에서 비주얼 특징을 추출한다. 추출된 비주얼 특징은 인식률을 높이기 위해 정규화 과정을 거친다. 정규화된 특징 값으로 인식 실험을 수행한다. 특징 추출과 입술 영상을 이용한 인식(립리딩)에 대해서는 Ⅵ장에서 상세히 설명하기로 한다.

음성 데이터 200개(10명 × 5모음 × 4회 발화)로부터 음성 특징(39차 MFCC: Mel frequency cepstral coefficient)을 추출하여 음성 인식 실험을 수행하였다. 음성 인식 실험의 상세한 사항은 Ⅴ장에 기술하였다. 음성 인식 결과와 립리딩의 통합은 음성 인식에서 최상위 신뢰도를 가진 인식 결과 후보와 두 번째로 신뢰도가 높은 후보와의 신뢰도 차이가 매우 적은 경우를 대상으로 한다. 이러한 경우 음성 인식 결과와 립리딩 결과를 통합하여 최종 인식 결과를 추정하는 신뢰도 재조정과정을 수행한다. 상세한 통합과정은 Ⅶ장에 기술하였다.

Ⅴ. 음성인식 실험

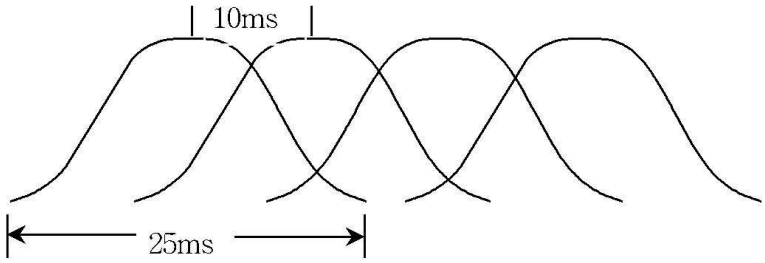
본 논문에서 음성 인식 실험은 HTK3.2 [9]를 사용하여 이루어졌다. HTK3.2에서는 HMM(Hidden Markov Model)을 사용한다. 다음은 인식 단위로 음소를 사용하는 경우 일반적으로 사용되는 HMM topology이다 [17].



[그림 9] HMM topology

[그림 9]의 HMM topology는 1 개의 Gaussian mixture와 5 가지 상태를 갖는 left-to-right 모델이다. 5개의 상태 중 첫 번째와 마지막 상태는 관측 심볼을 소비하지 않은 비출력(non-emitting state) 상태이다. 이들은 HMM과 HMM 사이의 null transition을 위해 사용된다. 음성 신호를 프레임(frame)단위로 분석하여 특징 벡터들을 만들어내면 이것이 곧 HMM의 관측 심볼이다. HMM의 각 상태는 그것과 연관된 출력 확률 분포 함수를 가지고 있어서, 그 상태에서 특정 관측 심볼(특징 벡터)이 나타날 확률을 계산할 수 있다. 이때 이 출력 확률 값이 높으면 그 모델이 표현하고 있는 인식 단위일 가능성이 높다는 것을 의미한다 [7].

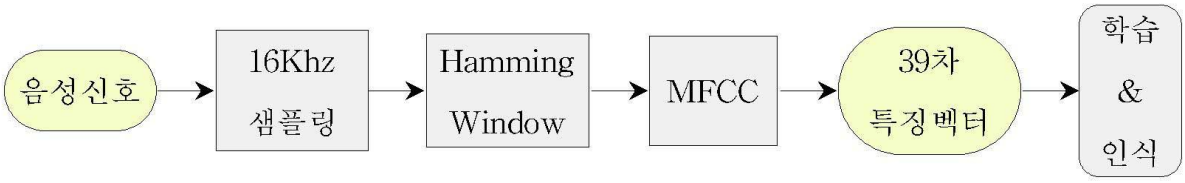
Baum-Welch 재예측 알고리즘(re-estimation algorithm)과 Viterbi 알고리즘을 사용하여 학습되고 테스트하였다. 음성 신호는 모두 16KHz로 샘플링 되었고, [그림 10]과 같이 매 10ms마다 25ms 크기의 해밍 윈도우를 만들어서 프레임을 분할하였다 [13].



[그림 10] Hamming Window

각 프레임에서 12개의 MFCC(Mel frequency cepstral coefficient) parameter와 첫 번째와 두 번째 순위 간 차의 계수로 구성된 39((12+1)*3) 차 특징 벡터를 추출한다.

[그림 11]은 본 실험의 음성 인식을 위한 데이터 전처리 과정을 나타낸다.



[그림 11] 음성 처리 과정

음성 인식 실험을 위한 HTK 수행 순서는 다음과 같다. 우선 각 모음에 맞춰 트랜스크립션(Transcription) 파일과 발음 사전을 만든다. 트랜스크립션 파일인 words.mlf와 phones0.mlf는 인식하고자 하는 다섯 모음들의 리스트를 나열한 파일과 인식하고자 하는 단어 리스트들에 대응하는 폰 열(phone list)을 나열한 파일이다. 다음 단계로 사람의 음성을 인식하기 위해서 음성의 특징들을 뽑아낸다. 일반적으로 LPC(Linear Prediction Coefficients)나 MFCC를 많이 사용하는데 본 논문에서는 MFCC를 사용하였다. 다음으로 음향 모델을 생성하고 이 모델을 반복적으로 학습시킨다. 마지막으로 발음사전과 인식 네트워크를 작성한 후 인식 실험을 한다.

[표 12]는 5개의 한글 모음에 대해 음성만을 인식한 결과이다. 각 행은 LLR(Log Likelyhood Ratio)에 기반한 N-best결과에 대응한다. 총 200개의 오디오 파일 중 160개의 데이터를 이용하여 학습하고 40개의 데이터로 테스트한 음성 인식 결과로 1-best일 때 65.0%의 인식률을 보였다.

N-best	Hit/#40	인식률
1	26	65.0%
2	32	80.0%
3	34	85.0%
4	39	97.5%
5	40	100%

[표 12] audio-only 인식 성능

일반적인 음성 인식의 정확도는 다음 식에 의해 계산된다.

$$Accuracy = \frac{N - D - I - S}{N} \times 100 \quad (1)$$

(단, N: 단어 수, D: 삭제 오류 개수, I: 삽입 오류 개수, S: 대체 오류 개수)

1-best 만 고려한 음성 인식 실험 결과는 [표 13]과 같이 대체 오류 개수가 14개로 정확도는 65%이다.

D	S	I	N	정확도
0	14	0	40	65%

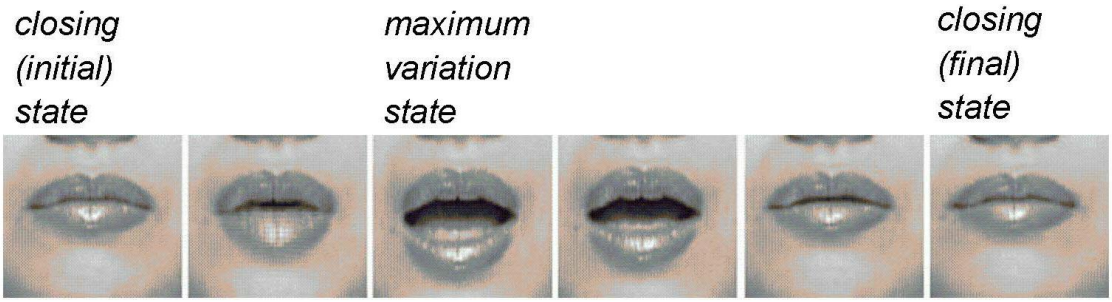
[표 13] 1-best 음성 인식 결과

Ⅵ. 립리딩 실험

1. 특징 추출

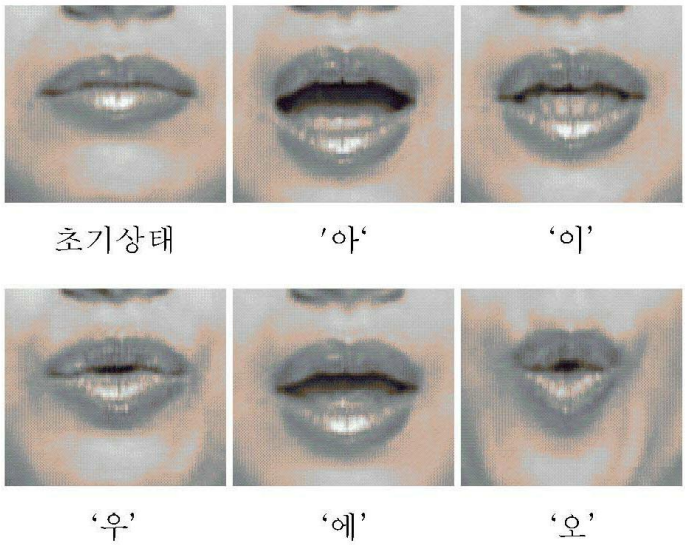
발화시 입술은 연속적인 변화를 보인다. 이 입술 움직임으로부터 특정 입술 이미지를 추출한다.

[그림 12]은 모음 '아'에 대한 입술 움직임의 연속이다. 각 모음의 발화시 입술은 다문 초기 상태(initial state)에서 시작하여 소리를 만들기 위해 최대변화 상태(maximum variation state)에 도달한 후 다시 다문 상태(final state)로 돌아간다. 이미지의 연속들로부터 특징 값 추출을 위해 각 모음의 초기 상태와 최대 변화 상태의 이미지를 선택하였다.

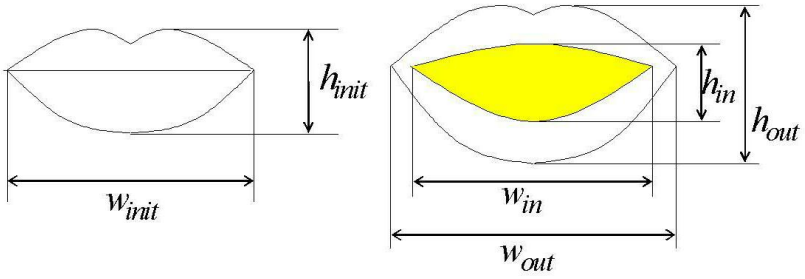


[그림 12] 발화시 입술 변화

[그림 13]은 입술을 다문 초기 상태와 5 가지 한국어 기본 모음 5 가지에 대한 최대 변화 상태의 정지 영상이다. 각 기본 모음에 대한 최대 변화와 초기 두 상태로부터 입술 높이와 너비에 대한 6 가지 간격을 [그림 14]와 같이 추출하였다.



[그림 13] 한국어 5 가지 기본 모음과 초기 상태



[그림 14] 6 가지 기하학적 간격

[그림 14]는 효과적인 입술 특징을 찾기 위한 사전 단계로 6 가지 기하학적인 간격을 나타낸다. 6 가지 간격은 w_{init}(초기 상태에서의 입술 너비), h_{init}(초기 상태에서의 입술 높이), w_{in}(입술을 벌렸을 시 안쪽 너비), h_{in}(입술을 벌렸을 시 안쪽 높이), w_{out}(입술을 벌렸을 시 바깥 너비), h_{out}(입술을 벌렸을 시 바깥 높이)이다.

그러나 각 화자 별로 입술형태나 크기 등이 다르고 모든 각도에서 동일한 크기로 촬영된다는 보장이 없기 때문에 이 6 가지 간격은 정확성이 떨어진다. 따라서 보다 효과적인 입술 특징을 위해 6 가지 기하학적 간격간의 다양

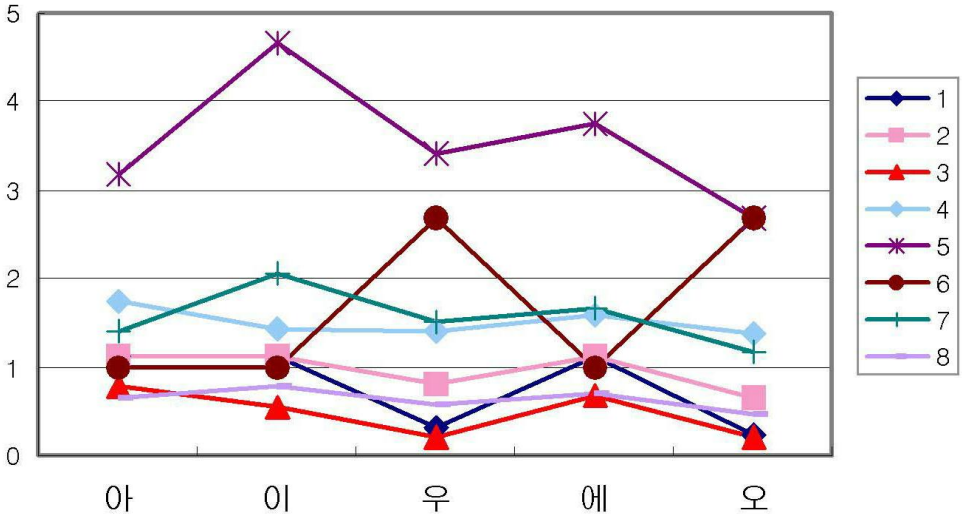
한 비율을 고려하였다. 그 중 모음간의 높은 식별력을 가지는 다음의 8 가지 비율을 특징 값으로 선택하였다 [8].

$$1)\ \frac{w_{in}}{w_{init}}, \qquad 2)\ \frac{w_{out}}{w_{init}}, \qquad 3)\ \frac{h_{in}}{h_{init}},$$

$$4)\ \frac{h_{out}}{h_{init}}, \qquad 5)\ \frac{w_{in}}{h_{in}}, \qquad 6)\ \frac{w_{out}}{w_{in}},$$

$$7)\ \frac{(w_{in}/h_{in})}{(w_{init}/h_{init})}, \qquad 8)\ \frac{(w_{out}/h_{in})}{(w_{init}/h_{init})}$$

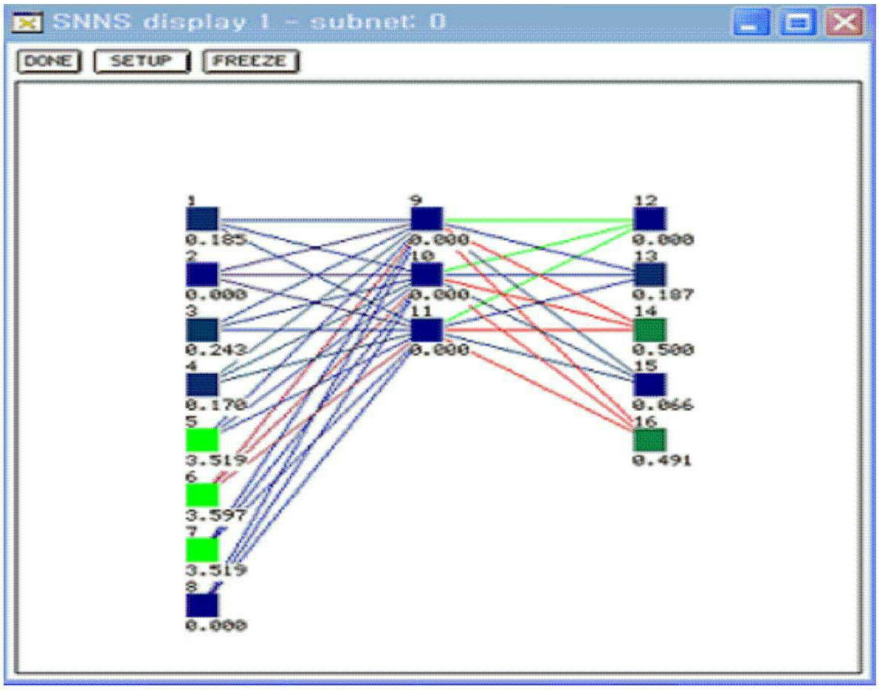
[그림 15]는 8 가지 특징 값에 대한 비교 그래프이다. 그래프를 보면 몇몇 특징 값(2번, 3번, 4번, 8번) 들이 변별력을 가지긴 하나, 변화 폭이 큰 다른 특징 값들에 비해 상대적으로 변이 폭이 적음을 알 수 있다. 이는 인식에 좋지 않은 영향을 미치는데 정규화 과정을 통해 보완 하였다.



[그림 15] 8가지 특징 값 비교 그래프

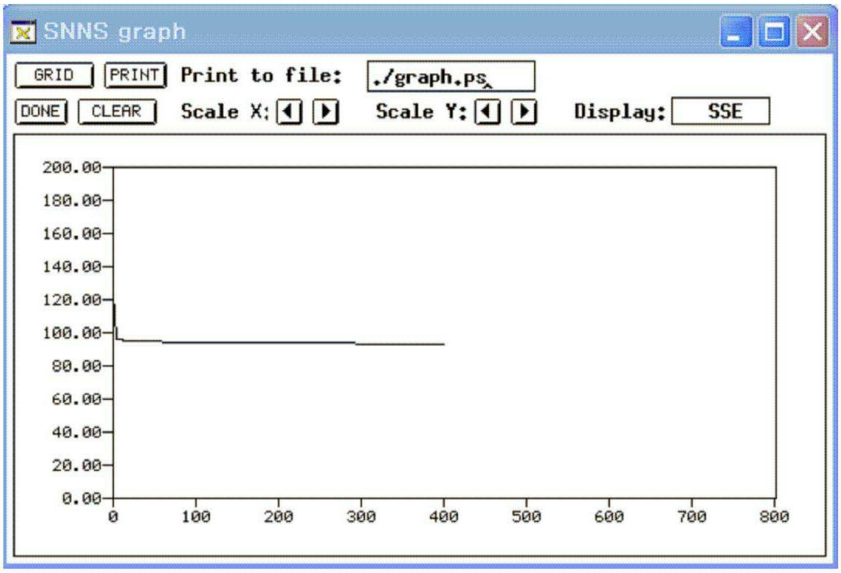
2. 정규화 및 인식

립리딩을 위해 SNNS v4.2 [10]를 사용해서 [그림 16]처럼 8개의 입력층, 3개의 은닉층 그리고 5개의 출력층으로 이루어진 back propagation 네트워크를 설계하였다 [10].

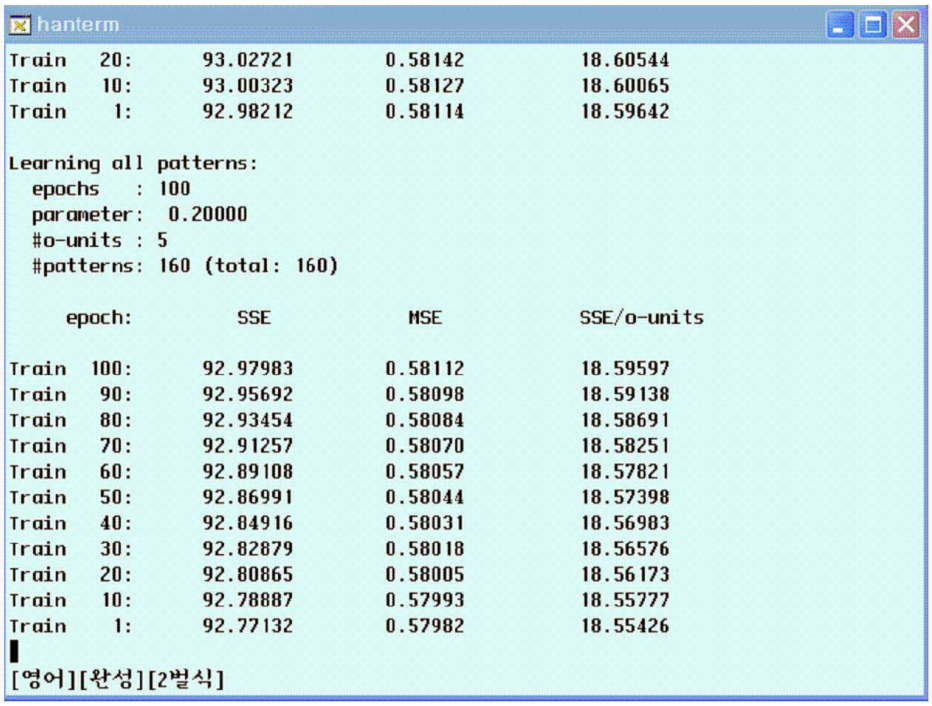


[그림 16] SNNS 설계 네트워크

설계된 네트워크를 [그림 17]의 그래프와 같이 일정한 값으로 수렴 할 때까지 학습 데이터를 사용하여 학습 시킨다.



[그림 17] SNNS 수렴 그래프



[그림 18] SNNS 학습과정

정면에 대한 총 200개의 비디오파일(10명 × 5모음 × 4번 발화) 중 160개는 학습 데이터로 40개는 테스트 데이터로 사용하였다. 입술 인식 성능은 테스트 데이터 40개중 22개가 올바로 인식되어 인식률이 [표 14]와 같이 55%로 나왔다.

[표 14]는 [그림 15]에서 보인 8개의 특징 값을 이용한 결과이다. [그림 15]에 나타난 바와 같이 8개의 특징 값 중 2번, 3번, 4번, 8번은 절대 크기가 다른 1번, 5번, 6번, 7번 보다 작아 개별적으로는 변별력이 있음에도 불구하고 인식 시 1번, 5번, 6번, 7번에 비해 상대적으로 변별력이 없는 특징으로 작용한다.

	아	이	우	에	오
아	**5**	0	0	3	0
이	0	**6**	0	0	2
우	0	0	**8**	0	0
에	0	5	0	**3**	0
오	0	0	8	0	**0**

[표 14] 정규화 전 인식률

이와 같이 변별력은 있으나 변이 폭이 다른 특징에 비해 상대적으로 작은 특징이 있으므로, 인식률을 보다 높여보고자 정규화를 시도하였다. 변이 폭을 일정하게 하기 위해 각 특징 값 별로 다섯 모음들을 식 (2)를 사용하여 0에서 5사이로 정규화 시켰다.

$$x_{normalized} = \frac{(value - \min) \times 5}{\max - \min} \qquad (2)$$

정규화 후 인식 실험 수행 결과([표 15])는 40개중 25개가 올바르게 인식되어 인식률이 55%에서 62.5%로 향상되었다. 세부적으로는 모음 '아'와 '에'의 인식률이 상승하였다.

[표 15]에서 모음 '오'와 모두 '우'로 오인식 되었다. 그러나 모음 '오'에 대한 '우'로의 신뢰도와 '오'로의 신뢰도 차이는 매우 미미하였다([표 17]의 visual 열의 입력 '오'에 대한 신뢰도 참조).

	아	이	우	에	오
아	**8**	0	0	0	0
이	0	**4**	0	2	2
우	0	0	**8**	0	0
에	3	0	0	**5**	0
오	0	0	8	0	**0**

[표 15] 정규화 후 인식률

3. 각도 별 립리딩 인식률 비교 분석

6 가지 각도 별로 립리딩 실험을 수행하였다. 각 각도 별로 200개의 이미지 데이터를 사용하였다. 이중 평균 5개의 에러 데이터가 있었고 155개를 학습 데이터로, 40개는 테스트데이터로 사용하였다. 인식 실험 수행 결과는 [표 16]과 같다.

수직 / 수평	-15°	0°	+15°
0°	70.0%	62.5%	57.5%
10°	-	67.5%	-
20°	-	47.5%	-
30°	-	37.5%	-

[표 16] 각도 별 립리딩 결과

수평 각도가 0°에서 멀어짐에 따라 인식률은 떨어진다. 단 수직 각도는 데이터 촬영 당시 입술 기준이 아닌 얼굴 미간을 기준으로 수직 각도를 측정하였으므로 -15°가 입술 정면에 더 가까운 각도가 되어 오히려 인식률이 더 높게 나타났다.

Ⅶ. 음성 기반의 립리딩 통합

본 논문에서는 이전 연구 [11][17][18]에 기반하여 음성 인식에서 첫 번째와 두 번째 후보자의 LLR 차가 임계치 보다 작은 경우에 영상 부분의 인식 결과를 사용하여 신뢰 점수의 N-best rescoring을 시도하였다. 음성 인식만의 N-best 결과는 Ⅴ장의 [표 12]와 같다.

1. LLR을 이용한 rescoring

음성인식과 립리딩 각각의 신뢰도를 이용한 N-best rescoring 과정은 다음과 같다.

1번째 후보자의 정규화된 LLR과 2번째 후보자의 차이가 클 때는 인식이 제대로 된 경우로 간주하고 likelihood 값이 거의 차이가 나지 않는 경우는 오인식으로 본다 [11][17][18]. 즉 LLR 간의 차가 특정 임계치 λ보다 작을 때만 대체오류 수정을 한다. 학습 데이터베이스 기반의 음성 인식 결과 분석을 통해 본 논문에서는 $\lambda = 2.072$ 로 하였다.

$$\frac{1}{T_{1st}} log P(O;M)_{1st} - \frac{1}{T_{2nd}} log P(O;M)_{2nd} < \lambda \qquad (3)$$

립리딩의 N-best score를 이용하여 음성 인식의 대체 오류 수정을 수행하였다.

음성 인식에서 k번째 후보자의 신뢰 점수 $as_{k,n}$는 우선, 음성 인식의 N-best LLRs를 (0, 1) 범위의 신뢰 점수로 재조정한다.

$$as_{k,n} = 1 - \frac{LLR_{k_{th}}}{\sum_{i=1}^{n} LLR_{i_{th}}} \qquad (4)$$

립리딩의 N-best 신뢰 점수를 사용한 N-best 음성 인식 결과에서 k번째 후보자의 재조정된 신뢰 점수는 다음 식과 같다.

$$\theta as \circ (1-\theta) vs_k \qquad (5)$$

여기서, vs_k는 신경망을 통해 얻은 립리딩에서 k번째 (음성에서) 후보자의 신뢰도 값이다. θ는 음성과 립리딩의 가중치로 본 연구에서는 0.5로 하였다.

[표 17]에서 음성 인식 결과와 입술 인식 결과를 통합한 예를 보이고 있다. '오디오 value'는 프레임당 음성의 LLR을 나타내며 '오디오 value'를 식 (4)로 음성 신뢰 점수로 변경한다. λ값이 2.072보다 적을 경우에만 립리딩의 신뢰 점수인 '비주얼 value'와 '오디오 value'를 식 (5)를 통해 재조정한다. 2번째까지의 후보자 중에서 식 (5)의 결과가 큰 값이 최종 인식 결과가 된다.

input	Audio				Visual		2-best rescoring	
	N-best	Value	Hit	1_{th}-2_{nd}	Value	Hit	Value	Hit
아	아	-48.905441	1	0.265682	0.703910	1	0.352908	1
	우	-49.171124			0.000000		0.000000	
	오	-50.243912			0.000000			
	이	-51.392876			0.203030			
	에	-52.559372			0.442290			
이	오	-48.363216	0	0.143120	0.007500	1	0.003756	1
	이	-48.506336			0.191980		0.095848	
	에	-48.507874			0.144690			
	우	-49.145954			0.002380			
	아	-53.382797			0.000010			
에	에	-47.905560	1	0.973209	0.418570	0	0.211389	1
	오	-48.878769			0.000000		0.000000	
	아	-49.926727			0.519430			
	우	-50.836967			0.000000			
	이	-51.446972			0.202330			
오	아	-42.378307	0	1.471096	0.000030	0	0.000015	1
	오	-43.849403			0.489460		0.240555	
	이	-44.099987			0.200430			
	우	-44.511612			0.493140			
	에	-45.525879			0.081270			

[표 17] 립리딩 결과와 음성 인식 결과의 rescoring 결과

Input '오'를 보면 음성에서 '아'로 잘못 인식된 값이 rescoring을 통해 '오'로 올바르게 인식되는 것을 알 수 있다. 또한 '오'와 '우'의 visual value 들이 미미한 차이를 보이는 것 역시 확인 할 수 있다. 이 때문에 비주얼 파트에서 '오'가 '우'로 전부 오 인식된다.

[표 17]의 첫 번째 input '아'는 오디오와 비주얼 각각이 모두 올바르게 인식됐을 때이며 2-best rescoring 결과도 올바로 인식된 경우이다. 두 번째 input '이'는 오디오는 올바르게 인식되지 못했으나 비주얼과 통합한 2-best rescoring은 결과가 올바르게 나온 경우이다. 마지막 input '오'는 음성과 비주얼 모두 1-best에서는 오 인식 됐으나 통합에 의한 rescoring 결과는 올바르게 나온 경우이다.

2. 각도에 따른 음성과의 통합 비교 분석

2.1 6가지 각도 별 오디오-비디오 음성 인식

각도별 립리딩 결과를 사용한 인식에서 일반적으로 정면의 경우가 가장 인식률이 높았고 수평이나 수직으로 기울수록 인식률이 떨어졌다. (단, 본 실험을 수행한 데이터베이스 수집에서 정면 각도를 맞출 시에 입술이 아닌 미간에 기준을 두었기 때문에 수직 각도는 이점을 고려하여 결과를 보아야 한다.) 립리딩의 인식 결과는 각도에 따라 심하게 영향을 받지만 음성 인식 결과와 결합을 하고나면 심한 편차가 많이 보완 된다.

[표 18]은 각도별 오디오-비주얼 통합 인식 결과를 나타낸다. 각도별로 통합 인식 결과를 정면(수평0° • 수직0°)의 인식률과 비교해 봤을 때, 수평과 수직 모두 ±15도 이내에서는 3% 이하의 인식률 저하를 보였다. 그러나 20도 이상의 각도 변화가 있을 때에는 22%, 6%와 같이 큰 편차를 보이며 인식률이 저하되었다.

수평	수직	인식률			인식률 저하도 (정면기준)
		립리딩	음성	결합 후	
0°	-15°	70.0%	65.0%	77.5%	0%
0°	0°	62.5%		**77.5%**	-
0°	+15°	57.5%		75.0%	-3%
10°	0°	67.5%		75.0%	-3%
20°	0°	47.5%		60.0%	-22%
30°	0°	37.5%		72.5%	-6%

[표 18] 각도 별 인식 결과 통합과 통합 후의 추이

2.2 15° 이하 각도에서의 오디오-비주얼 음성 인식

본 절에서는 [표 18]에서와 같이 높은 인식률을 가지는 15°이하의 데이터를 통합하여 전체적인 인식 실험을 하였다. 4 가지 각도에 대한 총 800개 (10명 × 5명 × 4번 발화 × 4가지 각도)의 데이터 중에서 에러 처리된 5개를 제외한 635개를 학습 데이터로 사용하고 나머지 160개를 테스트 하였다.

4 가지 각도의 경우를 모두 통합한 전체적인 오디오-비주얼 인식 결과는 [표 19]와 같다. 립리딩 인식률은 60%로 각 각도 별로 인식했을 때보다는 떨어졌으나 음성과의 통합과정을 거치면 76%로 큰 폭으로 보완되었다.

	음성	립리딩	통합
인식률	65.0%	60.0%	76.0%

[표 19] 4 가지 각도 통합 인식 결과

Ⅷ. 결론 및 향후 연구 과제

1. 결론

본 논문에서는 PDA등과 같이 제한된 컴퓨팅 파워를 가진 이동형 시스템에서 쉽게 추출할 수 있는 단순 입술 특징을 제안하였다. 그리고 이러한 단순 입술 영상 특징에 기반한 오디오-비주얼 음성 인식 방법을 제안하고 다양한 영상 촬영 각도에서 그 인식 성능을 실험하였다. 오디오만을 이용한 음성 인식의 첫 번째 후보자와 두 번째 후보자 사이의 LLR(Log likelihood Ratio) 차이가 임계 값 보다 적은 차이를 보이는 것을 대상으로 영상 부분 립리딩의 N-best 결과를 사용하여 대체 오류 수정(N-best rescoring)을 시도하였다. 실험의 결과는 제한된 입술 특징에 기반하여 본 논문에서 제안된 대체 오류 수정 방법이 WER(Word Error Rate)을 줄이는데 매우 효과적이라는 것을 보여준다. 특히 수직과 수평 모두 15° 이내의 입술 정보 값이 인식률을 높이는데 매우 효과적이었다. 제한된 입술 특징을 사용한 음성 인식 결과는 각 각도 별로 인식했을 때와 4가지 각도를 통합하여 인식했을 경우 모두 65%에서 평균 76%로 향상되었다. 이는 음성만으로 인식했을 때보다 상대적으로 17% 정도 인식 정확도가 증가한 것이다.

2. 향후 연구 과제

오디오-비주얼 데이터베이스를 처리하는 데는 영상과 음성을 각각 처리하고 통합하는 과정이 필요하다. 영상 처리의 경우 얼굴을 추적을 한 후, 입술 부위로 범위를 국한한 후 영상 특징을 추출하는 여러 단계의 과정이 필요하다. 음성 특징 추출은 기존의 연구 결과로도 별 무리 없이 검출이 가능하다. 그러나 Face Tracking, Lip localization, Visual-features Extracion같은 영상 처리는 아직 각 분야별로 세부 알고리즘 개발이 필요하다. 또한 세부 알고리즘과 이를 활용한 인식 시스템 개발을 위해서는 충분한 학습 데이터가 추가로 구축 되어야 할 것이다.

또한 이렇게 개발된 오디오-비주얼 음성 인식 시스템을 PDA나 자동차 같은 특정 환경에 실용화 했을 시, 특화된 운용 환경에 따라 시스템 튜닝 작업이 다시 필요하다. 이 역시 대량의 학습 데이터가 요구된다. 그러나 오디오-비주얼 데이터를 대량으로 수집하는 데는 많은 시간과 인력, 비용이 든다. 이를 위해서는 대량의 오디오-비주얼 기반 음성 DB를 효과적으로 수집하는 프로세스에 대한 연구도 필요한 분야이다.

참고문헌 및 사이트

[1] IBM Corporation, "Audio Visual Speech Recognition," http://www.research.ibm.com/AVSTG/srec.html, 2004.

[2] AMP Lab, "Audio-Visual Speech Processing," http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing, 2003.

[3] Intel Corporation, "Visual Interactivity: Audio-Visual Speech Recognition," http://www.intel.com/research/mrl/research/avcsr.htm, 2003.

[4] 민소희, 김진영, 최승호, "입술정보를 이용한 음성특징 파라미터 추정 및 음성인식 성능향상," 말소리, No.44, pp. 83-91, 2002.

[5] J.-Y. Kim, J.-H. Lee, K. Shirai, "Development of a Lip-Sync Algorithm Based on an Audio-Visual Corpus," IEICE Transactions on Information and Systems, Vol.E86-D, No.2, pp. 334-339, 2003.

[6] K. Murai, S. Nakamura, "Face-to-Talk: Audio-Visual Speech Detection for Robust Speech Recognition in Noisy Environment," IEICE Transactions on Information and Systems, Vol.E86-D, No.3, pp. 505-513, 2003.

[7] 안동훈, "표제어간 음운 변화 현상의 동적 모델링을 이용한 한국어 연속 음성 인식," 석사학위 논문, 1998.

[8] J.-Y. Suh, K.-N. Lee, K.-H. Hong and Y.-J. Lee, "Correcting Korean Vowel Speech Recognition Errors with Limited Lip Features," in Proc. of ICSLP'2004, Vol. 3, pp. 2529-2532, 2004.

[9] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (v 3.2)*, Entropic Cambridge Research Laboratory, 2002.

[10] A. Zell, et al., *SNNS (Stuttgart Neural Network Simulator) Version 4.2,* Institute for Parallel and Distributed High-Performance Systems (IPVR), University of Stuttgart, 2000.

[11] 정두경, 김형순, "한국어 연결숫자 인식에서의 발화검증과 대체오류수정," 말소리, No. 45, pp. 79-92, 2003.

[12] 서재영, 이경님, 홍기형, 이용주, "모음 인식을 위한 립리딩의 촬영 각도에 따른 비교연구," 대한음성학회 추계 학술대회 논문집, pp. 139-142, 2004.

[13] M. Ravishankar, *Efficient Algorithms for Speech Recognition*, CMS-CS-96-143, 1996.

[14] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview", *in Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[15] C. Neti, G. Potamianos, J. Luettin, I. Mattews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, *Audio-visual Speech Recognition*, The Center for Language and Speech Processing Workshop 2000 Final Report, The Johns Hopkins University, 2000.

[16] D. Gibbon, I. Mertins, R. Moore, *Handbook of Multimodal and Spoken Dialogue System*, Kluwer Academic Publishers, 2000.

[17] M. Rahim, C.-H. Lee and B.-H. Juang, "Discriminativ

Utterance Verification for Connected Digits Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, pp. 266-277, 1997.

[18] R. Setlur, A. Sukkar and J. Jacob, "Correcting recognition errors via discriminative utterance verification," in Proc. of ICSLP'96, Vol. 2, pp. 602-605, 1996.

ABSTRACT

A Study on Audio-Visual Korean Vowel Recognition based on Simple Lip Features for Different Camera Angles

Suh Jae-Young

Department of Computer Science

Graduate School of SungShin Women's University

In this paper, we describe audio-visual Korean vowel recognition experiments by using a limited set of lip features. We propose the lip features extracted from a snapshot image, for each vowel speech, when the speaker's mouth reaches maximum variation compared with its closing state. The proposed lip features can be obtained in a simple and cost effective way. It is very useful for the devices having limited computing power such as PDA or smart phone. We also develop an N-best rescoring method to correct Korean vowel speech recognition errors by using visual lip recognition results as the supplementary information. Finally, we perform experiments by using the extracted lip features from various different video-taking angles. For the experiments, we construct an audio-visual database from 10 subjects(speakers). Using the database, we evaluate the developed rescoring method. In

the experiments, we use HTK3.2 for speech recognition and SNNS4.2 for lip-reading based on the proposed lip features. The experimental results show that the developed rescoring method are very effective on audio-visual Korean vowel speech recognition.

감사의 글

2년여 간의 대학원 생활을 어느덧 마감하게 되었습니다. 대학 4학년 때부터의 근 3년 가까이 부족한 저를 보살펴 주신 홍기형 교수님과 유원경 교수님께 진심으로 감사드립니다. 바쁘신 가운데에도 논문 심사를 맡아주신 김호성 교수님께도 깊은 감사를 드립니다. 또한 대학과 대학원 생활을 포함한 지난 6년간 많은 가르침을 주신 박문화 교수님, 박종수 교수님, 우종정 교수님, 김도형 교수님, 심광섭 교수님, 서동수 교수님, 홍의석 교수님, 이재원 교수님, 서경희 교수님, 유민호 교수님, 장선희 교수님, 진경아 교수님께도 진심으로 감사드립니다.

처음 연구실 생활을 시작할 때 많은 조언을 주신 권하정 선배, 박아영 선배, 안미숙 선배, 이진숙 선배, 이하정 선배, 천기숙 선배 그리고 장재경 선배에게 감사드립니다. 논문 쓰는 마지막 순간까지 함께 고생해준 대학 동기이자 연구실 식구인 지혜와 영희에게도 감사의 마음을 전합니다. 그리고 이경님 선생님, 이경아 선생님, 연구실의 막내 지영이, 같은 연구실은 아니지만 항상 즐거움을 같이 나눈 문정 언니, 선배인 지선언니, 은정 언니, 세화 언니, 대학원 동기인 계숙 언니, 은경 언니, 미영에게도 감사의 마음을 전합니다. 그 외 못다 적은 많은 선배, 동기, 후배들에게도 감사하다는 말을 전하고 싶습니다.

마지막으로 항상 곁에서 지켜봐 주신 부모께 감사의 말을 전하며 계속 발전하는 모습을 보여드리겠습니다.