

PCA와 HMM을 이용한 실시간 립리딩 시스템의 설계 및 구현

이지근[†], 이은숙^{**}, 정성태^{***}, 이상설^{****}

요 약

립리딩은 잡음 환경에서 저하되는 음성 인식률의 보상과 음성을 청취하기 어려운 상황에서의 음성인식 보조 수단으로 많은 연구가 시도되고 있다. 기존 립리딩 시스템은 인위적인 조명 환경이나 얼굴과 입술 추출을 위하여 미리 정해진 제한된 조건에서 실험되어 왔다. 본 논문에서는 화자의 움직임이 허용되고 컬러나 조명과 같은 환경 변화에 제한조건을 완화한 영상을 이용하여 실시간 립리딩 시스템을 구현하였다. 본 논문의 시스템은 범용으로 사용되는 PC 카메라를 통해 입력받은 영상에서 얼굴과 입술 영역을 실시간으로 검출한 후, 립리딩에 필요한 영상정보를 추출하고 이 입술 영상 정보를 이용하여 실시간으로 발성 단어를 인식할 수 있다. 얼굴과 입술 영역 검출을 위하여 조명환경에 독립성을 갖는 색도 히스토그램 모델을 이용하였고 움직이는 화자의 얼굴 추적을 위하여 평균 이동 알고리즘을 이용하였다. 검출된 입술 영역에서 학습과 인식에 필요한 영상 정보를 추출하기 위하여 PCA(Principal Component Analysis)를 사용하였고, 인식 알고리즘으로는 HMM을 이용하였다. 실험 결과 화자종속일 경우 90%의 인식률을 보였으며 잡음이 있는 음성과 합병하여 인식을 실험을 해 본 결과, 음성 잡음비에 따라서 음성 인식률을 약 40~85%까지 향상시킬 수 있었다.

Design and Implementation of a Real-Time Lipreading System Using PCA & HMM

chi-geun Lee[†], eun-suk Lee^{**}, sung-tae Jung^{***}, sang-seol Lee^{****}

ABSTRACT

A lot of lipreading system has been proposed to compensate the rate of speech recognition dropped in a noisy environment. Previous lipreading systems work on some specific conditions such as artificial lighting and predefined background color. In this paper, we propose a real-time lipreading system which allows the motion of a speaker and relaxes the restriction on the condition for color and lighting. The proposed system extracts face and lip region from input video sequence captured with a common PC camera and essential visual information in real-time. It recognizes utterance words by using the visual information in real-time. It uses the hue histogram model to extract face and lip region. It uses mean shift algorithm to track the face of a moving speaker. It uses PCA(Principal Component Analysis) to extract the visual information for learning and testing. Also, it uses HMM(Hidden Markov Model) as a recognition algorithm. The experimental results show that our system could get the recognition rate of 90% in case of speaker dependent lipreading and increase the rate of speech recognition up to 40~85% according to the noise level when it is combined with audio speech recognition.

Key words: Lipreading(립리딩), PCA(주성분분석), HMM(히든마코브모델), HCI(인간-컴퓨터 상호작용), Pattern Recognition(패턴인식), Computer Vision(컴퓨터 비전)

※ 교신저자(Corresponding Author) : 이지근, 주소 : 전북 익산시 신용동 344-2(570-749), 전화 : 063) 850-6886, FAX : 063)856-8009, E-mail : lcg74@wonkwang.ac.kr

접수일 : 2004년 6월 8일, 완료일 : 2004년 10월 11일

[†] 준회원, 원광대학교 대학원 컴퓨터공학과 박사과정

^{**} 준회원, (주) 뷰콤 개발팀 연구원

(E-mail : enlion@mucom.co.kr)

^{***} 종신회원, 원광대학교 전기전자및정보공학부 교수
(E-mail : stjung@wonkwang.ac.kr)

^{****} 정회원, 원광대학교 전기전자및정보공학부 교수
(E-mail : slee@wonkwang.ac.kr)

※ 본 연구는 2003년도 원광대학교의 교비 지원에 의해서 수행됨.

1. 서 론

사람과 컴퓨터의 인터페이스에서 음성을 사용하고자 하는 연구가 아주 오래 전부터 수행되어 왔고 많은 음성 인식 시스템들이 개발되었다. 그러나, 실제 환경에서 음성 인식이 보편적인 인터페이스 수단으로 활용되는데 있어서 잡음이 큰 문제로 대두되었다. 근래에는 잡음에 덜 민감한 음성 인식 시스템을 개발하기 위한 연구가 널리 수행되고 있는데, 비주얼 정보를 보조 수단으로 이용함으로써 음성 인식의 정확도를 높이려는 오디오-비주얼 음성 인식에 대한 연구가 매우 활성화되어 있다. 립리딩은 음성인식 분야 중 잡음 환경에서 현저하게 떨어지는 인식률을 높이기 위한 보상 방법의 하나로 화자의 입술을 포함한 영상 정보는 발성의 조음 현상을 반영하고 있기 때문에 잡음이 있는 음성 파라미터를 보상하는데 사용되고 있다. 이러한 이유로 국내외에서 립리딩 시스템에 대한 연구가 활발히 진행되고 있다. 그러나 기존 대부분의 립리딩 시스템에서는 화자의 움직임이 없고 피부색과 다른 단일 배경색을 갖는 입력 영상을 사용하거나 조명의 변화에 대한 인식을 변화를 줄이기 위해서 실험 영상의 조명을 인위적으로 조정하였고, 그림 1과 같이 입술 영역만을 입력 영상으로 사용하기도 하였다[1-3]. 또한 입술에 색칠을 하거나 마커를 부착하여 입술 영역을 미리 학습시켜 입술 경계선을 추출하기도 하였으며[1], 인식에 필요한 입술 영역 이미지 벡터를 수작업으로 선정하거나[2] 이미 최적의 환경에서 촬영된 입술 이미지 데이터베이스를 이용[3]하여 실험하는 등, 여러 가지 제한 조건을 전제로 하여 특정한 환경에서만 사용할 수 있게 하였다. 또한 기존의 립리딩 시스템은 대부분 고성능 카메라를 이용하였으며 서로 다른 시스템들끼리 오프라인 상태에서 실험데이터를 입력받아 학습과 인식을 함으로써 독립된 하나의 시스템을 이용한 실시간

학습과 인식을 지원하지 않는 경우가 많았다.[1-5]

따라서, 본 논문에서는 화자의 움직임이 허용되고 컬러나 조명과 같은 환경 변화에 대한 제한을 완화시킨 영상을 이용하였으며, 기존의 오프라인으로 실험되었던 립리딩 시스템에 반하여 립리딩 실험의 전반적인 과정인 얼굴영역 검출단계에서 인식단계까지의 과정이 한번에 이루어 질 수 있는 실시간 독립 립리딩 시스템을 구현하였다. 본 논문에서 구현한 시스템은 화자가 직접 자신의 학습 데이터를 시스템 온라인 상태에서 생성하고 추가할 수 있으며, 인식 데이터를 생성, 입력하여 시스템 온라인 상태에서 학습과 동시에 실시간으로 인식 실험을 할 수 있는 립리딩 시스템을 구현하는 데에 연구 목적을 두었다. 또한 본 논문에서 구현된 립리딩 시스템에 대하여 잡음이 있는 환경에서 저하되는 음성인식률의 보상 정도와 가능성을 실험하였고 구현된 립리딩 시스템만의 자체 인식률을 평가하였다. 또한 기존의 립리딩 시스템들이 대부분 고성능의 비디오 카메라를 사용하여 영상을 획득했던 것에 대하여 이를 범용으로 사용되는 저가의 PC 카메라로 대신하였으며, 입술영역 검출의 전단계인 얼굴영역 검출에서부터 립리딩 인식 실험까지의 과정이 한번에 순차적으로 이루어 지도록 시스템을 구성하였다. 구현된 립리딩 시스템은 얼굴과 입술 영역 검출을 위하여 색도(Hue) 히스토그램 모델을 이용하였다. 색도(Hue)는 RGB 컬러가 조명의 영향에 민감하다는 단점을 보완할 수 있기 때문에 영상 인식에서 많이 사용되는 컬러 모델이기도 하다. 따라서 얼굴과 입술 영역 검출을 위하여 입력 영상 안에서 픽셀들에 대한 색도 값의 분포를 나타내는 색도 히스토그램 모델을 이용하였다. 먼저 RGB 값을 갖는 입력 영상에서 얼굴이 위치해 있을 가능성이 있는 영역을 얼굴의 색도 확률분포 영역으로 정의하고, 이 영역을 구하기 위하여 입력 영상의 첫 프레임에서 얼굴색의 HSI 색도(Hue) 히스토그램 모델을 만든다. 이어서, 영상획득(Image capture)이 시작된 후 인식을 위해 입력되는 RGB 영상을 HSI 컬러로 변환한다. HSI 컬러로 변환된 영상에 이미 만들어진 얼굴 색도 히스토그램 모델을 적용하여 얼굴의 색도 확률 분포 영역을 검출하였다. 또한 이와 함께 평균 이동 알고리즘(Mean shift Algorithm)을 이용하여 얼굴 검출과 움직이는 화자를 추적하는데 사용하였다. 검출된 얼굴 영역에서는 입술 검출을 위

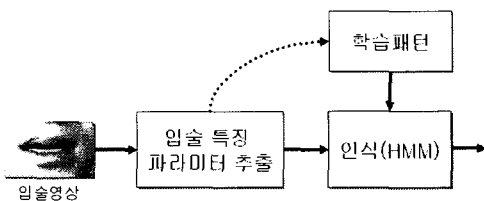


그림 1. 입술 영상만을 이용한 립리딩 시스템

하여 색도와 명도를 이용하였고, 검출된 입술 영역에서 학습과 인식에 필요한 영상 정보를 추출하기 위하여 PCA(Principal Component Analysis)를 사용하였다. 인식 알고리즘으로는 HMM(Hidden Markov Model)을 이용하였고, HTK(Hidden markov Tool Kit)를 사용하여 학습과 테스트 실험을 하였다. PCA를 통하여 추출된 특징 계수는 HTK에 적용할 수 있도록 데이터의 형식을 변형하여 사용하였다. 본 논문에서 구현한 립리딩 시스템을 이용하여 잡음이 있는 음성과 합병하여 인식을 실험을 해 본 결과, 음성 잡음비에 따라서 음성 인식률을 약 40~85%까지 향상시킬 수 있었고, 립리딩 시스템만의 자체 인식 실험 결과 화자 종속일 경우 90%의 인식률을 보였다.

2. 실시간 립리딩 시스템

2.1 시스템의 구성

본 논문에서 제안하는 립리딩 시스템의 전체적인 구성과 절차는 그림 2와 같이 크게 두 과정으로 나눌 수 있다. 첫 번째 과정은 학습데이터의 입술 영역 검출과 HMM을 이용한 학습이고, 두 번째 과정은 테스트 데이터의 입술 영역 검출과 인식이다.

학습데이터의 입술 영역의 검출과 HMM을 이용한 학습과정은 다음과 같다.

- 1) PC카메라로 320×240 크기의 영상 획득한다.
- 2) 입력 영상에서 일정 영역 내의 얼굴색의 픽셀로 색도 히스토그램 모델을 만들어 이를 다시 색도 영상에 적용하여 얼굴의 색도 확률 분포 영역을 구한다.

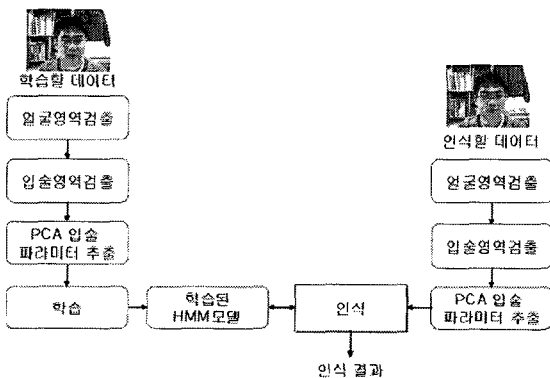


그림 2. 제안된 립리딩 시스템의 전체적인 구성도

- 3) 얼굴 확률 분포 영역에서 평균 이동 알고리즘 (Mean shift algorithm)을 이용하여 얼굴 영역의 검출과 추적을 수행한다.
- 4) 검출된 얼굴 영역에서 색도와 명도를 이용하여 입술 영역을 검출한다.
- 5) 검출된 입술 데이터에 PCA(Principal Component Analysis)를 적용하여 입술 영상의 계수를 추출한다.
- 6) 추출된 입술 영상의 계수를 HMM을 이용하여 학습시킨다.

테스트 데이터의 입술 영역 검출과 인식 실험 과정은 다음과 같다.

- 1) PC카메라로 320×240 크기의 영상을 입력받는다.
- 2) 학습데이터 준비과정과 동일한 방법으로 얼굴과 입술 영역을 검출하고 PCA 기법을 이용하여 입술 영역의 특징 계수를 추출한다.
- 3) 추출된 입술 영상 계수로 학습된 입술 데이터와 HMM 기법을 이용하여 입술 영상을 인식한다.

2.2 얼굴 영역 검출

영상획득 장치에서 획득된 컬러 영상은 RGB 컬러인데, 이는 조명 변화에 민감하게 반응하는 특성을 가지고 있다. 본 논문에서는 RGB 컬러를 HSI 컬러로 변환하여 원래 객체가 가지고 있는 색을 표현하는 색도 성분을 이용하여 영상이 조명에 영향을 적게 받도록 하였다. 얼굴 영역 검출에서는 입력 영상의 일정한 크기의 영역에서 픽셀들의 색도를 이용하여 색도 히스토그램 모델을 만들고 이를 색도 영상에 적용하여 얼굴의 색도 확률 분포 영역을 추출하였다. 이와 함께 얼굴 영역의 검출과 움직이는 화자의 얼굴을 추적하기 위한 방법으로는 평균 이동 알고리즘을 사용하였다.

2.2.1 얼굴 영역의 색도 확률 분포 계산

먼저, 얼굴 확률 분포 영역을 구하기 위하여 그림 3과 같이 색도 히스토그램 모델을 만든다. 입력영상의 첫 프레임에서 영상 중앙에 위치하는 64×64 크기의 윈도우를 지정하고 촬영 시작 전에 얼굴을 윈도우 크기보다 조금 더 크게 위치시킨다. 영상 입력 시에는 영상의 첫 프레임에서 윈도우 영역 안의 색도 값

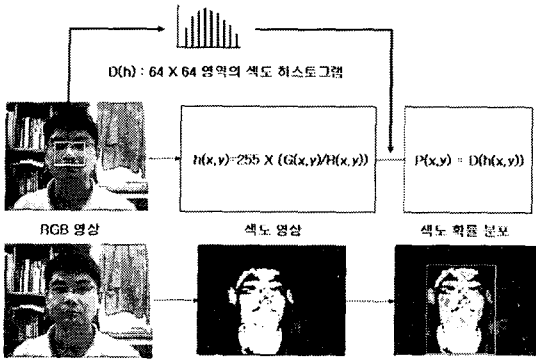


그림 3. 얼굴의 색도 확률 분포 추출 과정

을 추출하여 색도 히스토그램 모델을 생성한다. 색도 값을 구하는 방법으로 식(1)을 사용하였다.

색도 영상의 각 픽셀을 색도 히스토그램 모델 ($D(h)$)에 적용하여 얼굴의 색도 확률 분포 ($P(x,y)$)를 구하게 된다.

$$\begin{cases} Hue = 255 \times \frac{G}{R} & , R \geq G \\ 0 & , R < G \end{cases} \quad (1)$$

2.2.2 얼굴 영역 검출 및 추적

얼굴의 색도 확률 분포 영상에서 평균 이동 알고리즘을 사용하여 움직이는 화자의 얼굴영역을 추적할 수 있게 하였다. 얼굴 영역 검출과 추적을 위한 방법은 그림 4와 같은 순서로 진행한다. 먼저 색도 확률 분포 이미지에서 평균 이동 알고리즘을 이용하여 얼굴의 색도 확률 분포 영역에서의 중심을 찾고 다음으로 찾고자 하는 영역으로 수렴할 때까지 반복한다.

그림 4에서 새로운 탐색 윈도우의 중심을 검출하는 과정은 식(3)[6,7]을 이용하여 계산할 수 있다. 만약, $P(x,y)$ 가 이미지 안의 x, y 위치에서의 색도 값이고, 이를 x, y 에서의 확률 분포라고 생각할 때, 초기 적률(moment)과 1차 적률은 식(2)와 같다.

$$M_{00} = \sum_x \sum_y P(x, y) \quad \langle \text{초기 적률(moment)} \rangle$$

$$M_{10} = \sum_x \sum_y xP(x, y), \quad M_{01} = \sum_x \sum_y yP(x, y)$$

⟨1차적률(moment)⟩ (2)

식(2)를 사용하여 식(3)과 같이 평균 픽셀 값을 갖는 탐색 윈도우의 중심 위치를 구한다.

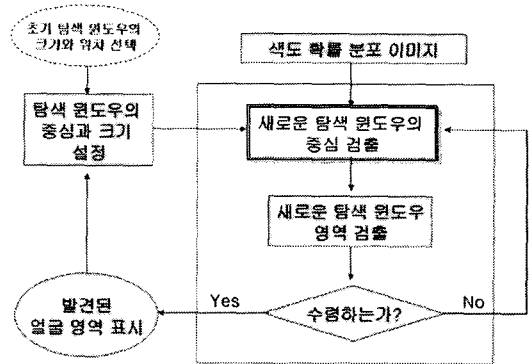


그림 4. 색도 확률 분포 이미지에서 얼굴 영역 검출 및 추적 순서도

$$x_c = \frac{M_{10}}{M_{00}} \quad y_c = \frac{M_{01}}{M_{00}} \quad (3)$$

새로운 탐색 윈도우 영역을 구하기 위해서는 먼저 식(4)와 같이 2차 적률(moment)를 구하고, 식(5)를 사용하여 탐색 윈도우의 방향을 구한다.

$$M_{20} = \sum_x \sum_y x^2 P(x, y) \quad M_{20} = \sum_x \sum_y y^2 P(x, y)$$

⟨2차 적률(moment)⟩ (4)

$$\theta = \frac{\arctan \left[\frac{2 \left[\frac{M_{11}}{M_{00}} - x_c y_c \right]}{\left[\frac{M_{20}}{M_{00}} - x_c^2 \right] - \left[\frac{M_{02}}{M_{00}} - y_c^2 \right]} \right]}{2} \quad (5)$$

탐색 윈도우의 방향을 구한 다음에는 식(6)과 식(7)을 이용하여 탐색 윈도우의 길이 (l)와 너비 (w)를 구한다.

$$l = \sqrt{\frac{(a+c) + \sqrt{b^2 + (a-c)^2}}{2}} \quad (6)$$

$$w = \sqrt{\frac{(a+c) - \sqrt{b^2 + (a-c)^2}}{2}} \quad (7)$$

여기에서, 중간 값 a, b, c 는 식(8)(9)(10)과 같다.

$$a = \frac{M_{20}}{M_{00}} - x_c^2 \quad (8)$$

$$b = 2 \left(\frac{M_{11}}{M_{00}} - x_c y_c \right) \quad (9)$$

$$c = \frac{M_{02}}{M_{00}} - y_c^2 \quad (10)$$

이와 같은 과정으로 새로 구해진 탐색 윈도우 영역이 이전의 탐색 윈도우 영역과 차이가 없다면, 임계값을 수렴한 것으로 간주하고 발견된 얼굴 영역을 표시한다. 이와 같이 현재 프레임에서 얼굴 영역을 발견한 다음에는 다음 프레임에서의 얼굴 영역 검출을 위해 탐색 윈도우를 새로이 설정한다. 연속된 프레임에서 얼굴의 크기가 변할 수 있으므로 식(11)을 이용하여 s 값을 구하고 탐색 윈도우의 너비는 $2s$, 길이는 $2.4s$ 로 설정하였다.

$$s = 2 * \sqrt{\frac{M_{00}}{255}} \quad (11)$$

그림 5의 (a)와 (b)는 색도 히스토그램 모델을 적용하여 얼굴의 색도 확률 분포를 구한 그림과 평균이동 알고리즘을 적용하여 얼굴을 추적한 그림이다. 배경에 얼굴의 색도 분포와 비슷한 색을 가진 영역이 있더라도 평균 이동 알고리즘을 적용함으로써 얼굴 객체만을 추출할 수 있다.

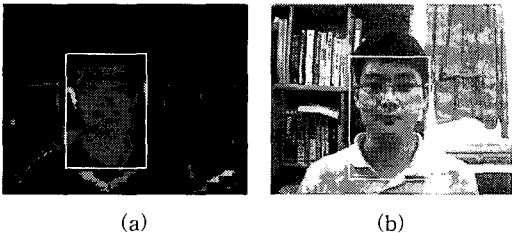


그림 5. (a) 얼굴의 색도 확률 분포 (b) 얼굴영역검출

2.3 입술 영역 검출

본 논문에서는 입술영역을 검출하기 위하여 색도 평균 마스크를 이용하였다. 입술은 얼굴의 크기에 비례적으로 있으므로 마스크의 크기를 얼굴 영역의 x 축의 $1/4$ 크기로 하였다. 색도 평균을 구하는 마스크로 얼굴 영역 내에서 회전 검색하여 입술이 분포한 영역을 찾는다. 입술은 얼굴의 안쪽 영역에 분포하므로 검색영역을 안쪽에서부터 시작한다. 입술의 R성분은 얼굴 영역의 다른 픽셀의 R성분 값보다 크다. 따라서, 입술의 색도 값은 식(1)에 의하여 작은 값의 분포를 갖게 된다. 즉, 입술 분포 영역 검출은 마스크의 색도 평균이 최저인 부분이 된다. 그림 6은 이와 같은 방법으로 검출된 입술영역이다.

개략적인 입술 영역을 검출한 후, PCA(Principle Component Analysis)를 이용하여 입술 파라미터를

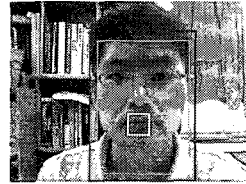


그림 6. 입술 영역 검출

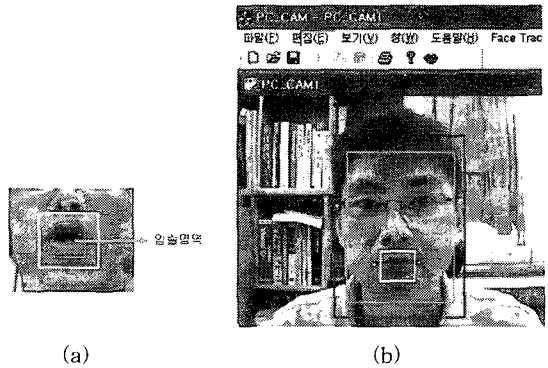


그림 7. (a) 입술 양 끝점 검출 영역 (b) 입술 양 끝점 검출

추출하기 위해 일정한 입술 위치와 크기를 갖는 입술 영역을 추출한다. 그림 7(b)와 같이 정확한 입술의 분할과 일정한 크기의 입술 영역을 추출하기 위해 입술의 양 끝점을 검출한다. 앞 단계에서 검출한 개략적인 입술 영역에서 입술의 양 끝점을 검출하기 위해서는 그림 7(a)와 같이 개략적인 입술 영역보다 더 큰 영역을 잡는다. 양 끝점의 검출은 입안의 명암도가 낮은 특성을 이용하는데 명암도는 식(12)와 같이 계산된다.

$$\text{명암도} = \frac{(R+G+B)}{3} \quad (12)$$

입술의 양 끝점을 추출하기 위해 개략적으로 검출된 입술 영역에서 식(12)로 각 픽셀의 명암도를 구하고 영역 내에서 각 열의 명암도 평균을 구한 다음 이 평균의 0.6 을 임계값으로 정하여 입술의 양 끝점을 추출한다. 추출한 입술의 양 끝점의 두 좌표는 입의 중앙 좌표를 결정하는데 이용되고 결정된 중앙 좌표를 기준으로 70×50 크기의 정규화 된 이미지로 저장한다. 검출된 입술 영역에서 PCA(Principle Component Analysis)를 이용하여 입술의 파라미터를 구하기 위해 그레이 레벨의 PGM(Portable Gray Map) 파일로 저장한다. 그림 8은 “다음”이라는 발음을 하였을 경우에 입력영상에서 추출되는 28 프레임의 입

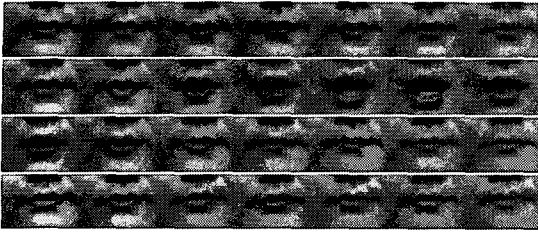


그림 8. 입술 영역 저장 프레임

술영역이다.

2.4 PCA를 이용한 입술 파라미터 추출

PCA(Principal Component Analysis)는 고차원 입력 벡터를 저차원의 벡터로 표현하여 몇 개의 주성분 값으로 나타내어 주는 방식이다. PCA를 이용하여 입술 영역 이미지의 특징 파라미터를 추출하는 것은 전체 영상의 데이터를 데이터의 분산이 큰 몇 개의 고유입술(EigenLips)에 대한 축으로 선형 투사시켜 데이터의 차원을 줄이는 것이다. 여기에서 고유입술은 원래의 입술 영역 이미지에 일치하는 공분산 행렬(Covariance Matrix)의 고유벡터(Eigenvectors)를 나타낸다. 그러므로 각 입술 영역 이미지의 주성분 값을 구하기 위해서는 공분산 행렬을 이용하여 고유벡터를 구해야 한다. 입술 이미지의 고유벡터를 구하는 과정은 다음과 같다.

입술 이미지의 학습 집합(X)를 M 개의 벡터로 표현하면 식(13)과 같고, 각 벡터는 N 개의 픽셀을 갖는다. 즉, 벡터 x 는 입술 이미지의 전체 픽셀 N 개를 갖는 벡터를 나타낸다.

$$X = \{x_1, x_2, \dots, x_M\} \tag{13}$$

M 개의 입술 이미지의 평균벡터 Ψ 는 식(14)와 같고,

$$\Psi = \frac{1}{M} \left[\sum_{i=1}^M x_i \right] \tag{14}$$

입술 이미지의 벡터 (x_i)와 입술 이미지의 평균 벡터 (Ψ) 사이의 차이는 식(15)와 같이 구할 수 있다.

$$\phi_i = x_i - \Psi, \quad i = 1, \dots, M \tag{15}$$

위의 식(15)를 이용하여 공분산 행렬 (C)을 구한다.

$$C = \frac{1}{M} \left[\sum_{n=1}^M \phi_n \phi_n^T \right] \tag{16}$$

행렬 (C)의 고유값 (λ)과 고유벡터 (e)는 식(17)의 방정식 해를 찾음으로써 구한다.

$$Ce = \lambda e \quad (x \neq 0) \tag{17}$$

계산된 고유벡터들의 행렬을 A 라고 하고, 다음 식(18)을 통해서 입술 이미지 벡터 x 의 주성분 값 (y)를 구한다.

$$y = A(x - \Psi) \tag{18}$$

식(18)을 이용하면 원래 입술 이미지 (x)로 복원하는 식(19)를 재구성할 수 있다.

$$x = A^T y + \Psi \tag{19}$$

본 논문에서는 행렬 A 의 고유벡터들을 모두 사용하지 않고 상위에서 K 번까지의 고유벡터를 사용하였다. 이를 다시 표현하면 식(20)과(21)과 같다.

$$y = A_K(x - \Psi) \tag{20}$$

$$x = A_{y+PK}^T \tag{21}$$

본 논문에서는 70×50 크기의 입술 이미지를 사용하였다. 학습에 사용된 입술 이미지가 총 739 프레임 일 경우, 학습 집합(X)의 각 벡터가 갖는 N 개의 픽셀은 3,500 픽셀이 된다. 그림 9는 학습 집합(X)를 나타낸다.

계산된 공분산 행렬 (C)는 739×739 의 행렬이 되고 공분산 행렬 (C)를 방정식(17)에 적용하여 고유값 (λ)과 739개의 고유벡터 (\bar{X})를 구한다.

본 논문에서는 전체 739개의 고유벡터 (\bar{X})중에서 표 1과 같이 상위 범위에 있는 고유벡터들을 각각 학습데이터와 인식 데이터의 각 프레임에 사영(projection)시켜서 추출되는 특징 계수들을 이용하

$$X = \begin{bmatrix} x_{1 \text{ frame}} \\ x_{2 \text{ frame}} \\ \vdots \\ \vdots \\ x_{739 \text{ frame}} \end{bmatrix} = \begin{bmatrix} x_{1.1} & x_{1.2} & \dots & x_{1.3500} \\ x_{2.1} & x_{2.2} & \dots & x_{2.3500} \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ x_{739.1} & x_{739.2} & \dots & x_{739.3500} \end{bmatrix}$$

그림 9. 학습 집합(X)

표 1. 특징계수에 따른 립리딩 인식률의 변화

| | | | | | | | | | |
|-----------------|------|------|------|------|------|------|------|------|-----|
| 상위 고유벡터범위(%) | 40 | 30 | 20 | 15 | 10 | 5 | 4 | 2 | 1 |
| 추출된 특징계수의 수 | 295 | 221 | 147 | 110 | 73 | 36 | 29 | 14 | 7 |
| 인식률(%) /10회 | 92.7 | 91.2 | 91.0 | 90.0 | 91.3 | 90.3 | 80.2 | 25.9 | 4.5 |

여 립리딩 실험을 10회 반복하였다. 화자중속일 경우, 추출된 특징 계수의 수에 따라서 다음 표 1과 같은 결과를 나타냈다.

전체 고유벡터 중에서 상위 1%~4% 내의 벡터에서 추출되는 특징계수의 수를 이용한 립리딩 인식 결과는 인식률 저하의 폭이 컸다. 따라서 고유벡터중 상위 5%의 벡터를 이용한 36개의 특징 계수를 립리딩 인식을 변화의 임계치 개수로 결정하였다. 그림 10은 원래 입술 이미지를 식(21)로 복원한 것이다.

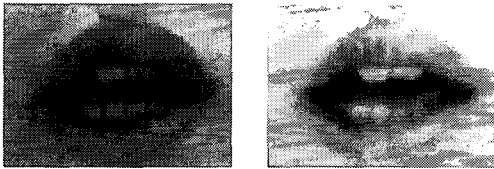


그림 10. (a) 본래의 입술 영상 (b) 고유벡터로 복원한 이미지

2.5 HMM을 이용한 립리딩

HMM은 다중확률 구조를 갖는 프로세스들을 모델링 하는데 매우 적합한 방법으로서 HMM 파라미터들의 정밀한 계산을 위한 효과적인 알고리즘이 존재한다는 장점 때문에 근래에 음성인식 뿐만 아니라 얼굴 인식, 제스처 인식, 영상 인식 등, 패턴인식 영역에서 전반적으로 널리 사용되고 있는 방법이다.

HMM은 확률의 상태와 그들 간의 천이 확률로 정의되며, 각 천이는 상태선택에 관한 상태 천이확률과 천이가 이루어졌을 때 유한개의 관측 심볼로부터 각 출력 심볼이 나올 수 있는 조건부 확률과 관련이 되어있다. 즉, HMM은 관측이 불가능한 하나의 프로세스를 관측이 가능한 심볼로 발생시키는 프로세스를 통하여 추정하는 이중확률 프로세스로서 음성이나 영상과 같이 가변성이 많고 발생과정을 알 수 없는 프로세스를 모델링 하는데 적합하다.

HMM은 다음과 같이 상태천이확률, 관측 심볼의 확률분포, 초기 상태확률의 3가지 확률 파라미터로 정의된다.

$$\lambda = (A, B, \Pi) \tag{22}$$

여기에서, A 는 상태천이 확률분포를 나타내는 상태천이 행렬이고, 다음과 같다.

$$A = [a_{ij}], \quad 1 \leq i \leq N \tag{23}$$

a_{ij} 는 상태 i 에서 상태 j 로 천이 될 확률이고 N 은 상태 수이다. B 는 관측 심볼의 발생확률 분포를 나타내며, 다음과 같다.

$$B = [b_j(k)], \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{24}$$

$$b_j(k) = P(o_t = v_k | q_t = j) \tag{25}$$

$b_j(k)$ 는 현재상태 q_t 에서 관측 심볼 (o_t)이 v_k 가 될 확률이고 M 은 관측 심볼의 개수이다.

마지막으로 Π 는 초기상태확률 분포이며, 다음과 같다.

$$\Pi = [\pi_1, \pi_2, \pi_3, \dots, \pi_N] \tag{26}$$

HMM은 forward-backward 알고리즘, Viterbi 알고리즘, Baum-Welch 재추정 알고리즘을 이용하여 인식문제에 적용되어진다. forward-backward 알고리즘은 관측 열 $O = [O_1, O_2, O_3, \dots, O_T]$ 와 HMM $\lambda = (A, B, \Pi)$ 가 주어졌을 때, 관측 열을 발생시킬 확률 $P(O|\lambda)$ 를 계산하는 확률 계산법이다. Viterbi 알고리즘은 최적의 상태 열과 그 상태 열을 통한 확률을 구하는데 사용되고, Baum-Welch 재추정 알고리즘은 초기 모델이 주어졌을 때, 학습 데이터를 사용하여 관측 심볼의 발생확률을 최대화하기 위해 HMM 파라미터 λ 를 반복적으로 학습시키는데 사용된다. 본 논문에서는 HMM 알고리즘이 적용된 HTK(Hidden Markov Toolkit)을 이용하여 학습과 인식을 하였다. HTK는 HMM인식 알고리즘을 수행할 수 있도록 HMM 모델 생성, 모델의 재추정과 학습 및 인식, 그리고 인식결과 분석에 필요한 라이브

러리를 지원하고 있는 인식관련 툴킷이다. 인식에 필요한 특징 데이터는 PCA에서 구한 특징 계수들을 HTK 파일 형식에 맞게 변환하여 이용하였고 사용된 HMM 모델은 3-state left-to-right 전이형태의 모델을 사용하였다. 그림 11과 12는 본 논문에서 실험한 HTK 진행 과정과 실험에 사용된 HTK 파일 형식이다.

| HTK 파일 형식 | |
|-----------|---|
| • 헤더 | •nSamples → 프레임 수 •sampSize → PCA에서 구한 특징 계수의 개수 X 4 •samPeriod → 333889(29.95Frames/sec일 경우) |
| 샘플 데이터 | |

그림 12. 논문에서 사용된 HTK 파일 형식

3 실험 결과

실험은 본 논문에서 구현된 립리딩 시스템을 사용하여 화자 독립일 때와 화자 종속일 때의 두 가지 경우에 잡음 환경에서의 음성인식을 저하를 보상하는 실험과 구현된 립리딩 시스템만의 자체적인 인식 실험을 하였다. 실험에 사용된 립리딩 학습데이터는 시스템 온라인 상태에서 저가의 범용 PC 카메라를 이용하여 영상을 입력받아 얼굴영역을 검출하고 검출된 얼굴 영역 내에서 입술영역을 검출한 후 실시간

으로 학습하였다. 학습데이터 단어로는 오디오 및 CD플레이어를 작동시키기 위해 필요한 단어들(재생, 정지, 종료, 앞으로, 뒤로)을 사용하였다. 입력되는 영상은 15 Frame/sec이고, 저장되는 입술정보는 대략 1020 프레임의 이미지를 얻었다. 입력 영상은 320×240 크기에 24비트 RGB 칼라 값을 갖는 프레임을 이용하였다. 구현 프로그램으로는 Visual C++ 6.0을 이용하여 얼굴영역 검출과 입술 영역 검출을 구현하였으며 학습 및 인식을 위하여 HTK를

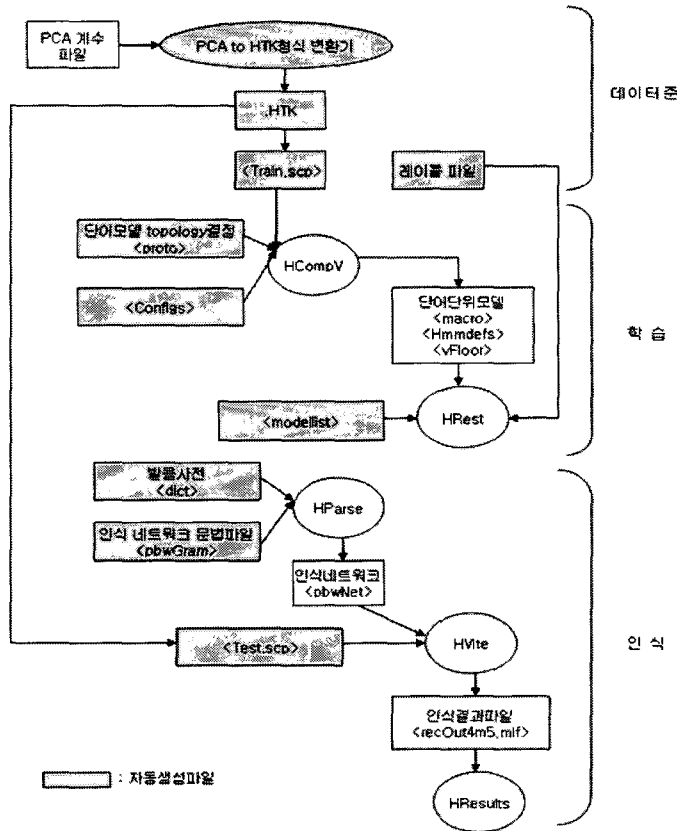


그림 11. HTK 인식과정

사용하였다.

3.1 잡음 환경에서의 인식을 보상

잡음 환경에서의 음성 인식을 저하를 보상하는 실험에서 사용된 립리딩 시스템의 전체적인 구조는 그림 13과 같다.

실험에서 사용된 음성인식 보완 립리딩 시스템은 그림 12와 같이 음성정보와 영상정보를 분석하여 특징 파라미터를 추출하고 학습과 인식을 한 다음 인식된 결과를 통합하게 된다.

3.1.1 화자독립 립리딩 시스템

첫 번째 인식 실험에서는 입술인식과 음성인식을 모두 화자독립으로 수행하였으며 학습데이터는 15명, 실험데이터는 2명으로 하였다. 음성인식의 경우, 5개의 단어를 연속 발음한 16비트 PCM형식의 AVI 포맷의 음성 데이터를 사용하였고, 이 음성 파일에 오디오 편집 프로그램인 쿨에디트 프로(Cool Edit Pro)를 이용하여 무 잡음의 음성 데이터에 Brown 잡음을 혼합하여 생성된 잡음-음성 데이터를 사용하였다. 음성 인식 알고리즘으로는 입술인식과 동일하게 HMM을 사용하였으며, 사용된 HMM 모델은 3 state left-to-right 전이형태의 HMM 모델을 사용하여 학습-인식 실험을 하였다. 학습과 인식 도구로는 입술인식과 음성인식 모두 HTK를 이용하였다. 실험 결과 화자독립 립리딩만의 인식률은 40%의 결과를 보였고, 음성만을 인식했을 경우 표 2와 같이 Brown 잡음의 비율에 따라 인식률이 변화하는 음성 데이터와 합병하여 실험하였다.

음성-입술을 합병하는 방법은 음성과 입술정보를

표 2. 잡음에 따른 음성 인식률의 변화

| 음성 잡음 | 5 | 10 | 20 | 30 |
|--------|-----|-----|-----|-----|
| 음성 인식률 | 80% | 35% | 30% | 20% |

각각 인식한 후 인식 결과를 합병하는 방법[8]을 사용하였다. 입술인식률(M_{lip})과 음성인식률(M_{speech}) 결과에 잡음도에 따른 가변적 가중치(α)를 주어 인식률을 합병하여 식(27)과 같이 합병 인식결과(M)를 구하였다.

$$M = \alpha M_{lip} + (1 - \alpha) M_{speech} \quad (27)$$

$$\alpha = \frac{3 \times \text{잡음도}}{100}, \quad 0 \leq \alpha \leq 0.9 \quad (\text{고정상수: } 3)$$

표 3은 입술인식, 음성인식 결과에 가중치를 달리 하여 합병한 후 인식한 결과이다.

표 3과 같이 0.1에서 0.9사이의 가변적인 가중치를 서로 다른 음성의 잡음도에 각각 적용하여 음성-입술 합병 인식률에 대한 실험을 하였다. 실험결과 값을 이용하여 각 음성의 잡음도에서 최적의 음성-합병 인식률을 결정하는 가중치를 얻을 수 있는 고정상수를 선택하였고, 각 잡음도에 고정상수 3을 곱하여 가중치를 계산하였다. 실험결과 잡음에 따른 가변적 가중치 α 에 의하여 음성-입술의 합병 인식률이 40~85%를 나타내는 것을 볼 수 있다. 그림 14는 본 실험의 입술, 음성, 음성-입술 합병 인식결과를 보여준다.

3.1.2 화자 종속 립리딩 시스템

화자종속 립리딩 시스템은 앞 절의 화자독립 립리딩 시스템과 동일한 방법과 조건으로 수행하였으며 학습데이터는 8회 발성한 영상 및 음성을 사용하였

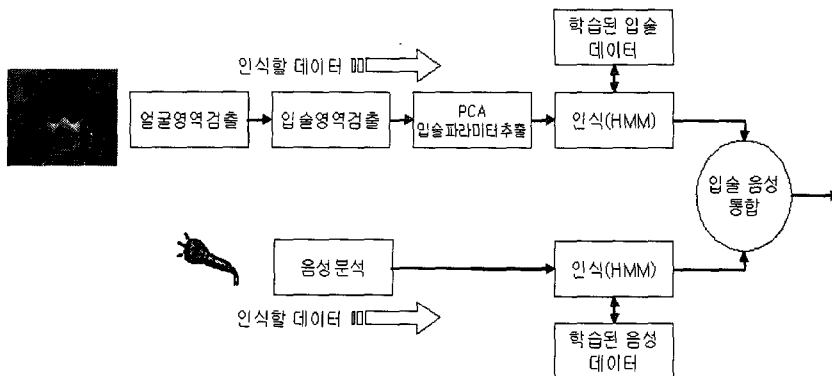


그림 13. 음성 인식 보상 립리딩 시스템의 전체적 구조

표 3. 가변적 가중치에 따른 합병 인식률

| 가중치(a) | | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---------------|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 음성입술합병 인식률 | 잡음 30 | 20 | 20 | 25 | 25 | 25 | 30 | 35 | 35 | 40 |
| | 잡음 20 | 30 | 30 | 30 | 30 | 30 | 25 | 25 | 35 | 40 |
| | 잡음 10 | 35 | 40 | 45 | 40 | 50 | 50 | 45 | 45 | 45 |
| | 잡음 5 | 80 | 85 | 85 | 75 | 75 | 75 | 55 | 55 | 45 |

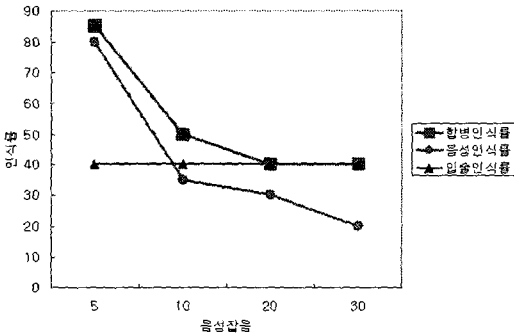


그림 14. 화자독립 입술, 음성, 음성-입술 합병 인식결과

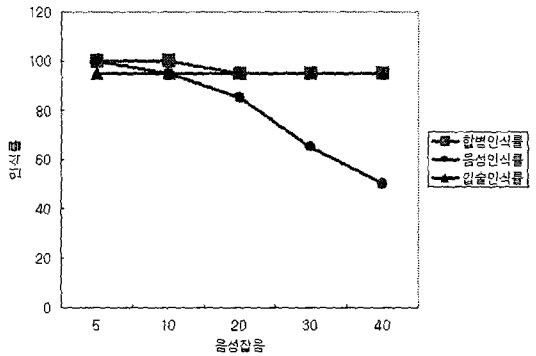


그림 15. 화자 종속 입술, 음성, 음성-입술 합병 인식결과

고, 실험데이터는 2회 발생한 영상과 음성으로 하였다. 화자종속 립리딩만의 실험결과 95%의 인식률을 보였고, 음성만을 인식했을 경우 표 4와 같이 Brown 잡음의 비율에 따라서 인식률이 변하는 데이터와 합병하여 실험하였다.

표 5는 입술인식과 음성인식 결과에 가변적 가중치 α 를 사용하여 합병한 인식 결과이다.

그림 15는 화자종속 립리딩 실험의 입술, 음성, 음성-입술 합병 인식결과를 보여준다.

표 4. 잡음 비율에 따른 음성 인식률의 변화

| 음성 잡음 | 5 | 10 | 20 | 30 | 40 |
|-------|------|-----|-----|-----|-----|
| 음성인식률 | 100% | 95% | 85% | 65% | 50% |

표 5. 화자종속입 경우 합병 인식률

| 음성잡음 | 5 | 10 | 20 | 30 | 40 |
|-------|------|------|-----|-----|-----|
| 합병인식률 | 100% | 100% | 95% | 95% | 95% |

3.2 입술 정보만을 이용한 립리딩 실험

본 논문에서 구현한 립리딩 시스템을 이용하여 입술 정보만을 활용한 실시간 인식 실험을 수행하였다. 먼저 사용자가 학습데이터 5개의 단어(재생, 정지, 종료, 앞으로, 뒤로)를 8회 발음한 영상을 준비한다.

그림 16은 논문에서 구현된 프로그램 상에서 PC카메라를 이용한 영상획득과 얼굴추적 및 입술검출을 시스템 온라인 상태에서 실행시킨 그림이다.

다음으로 그림 15에서 검출된 입술 영역에 대하여 그림 16과 같이 저장될 입술 영상의 파일 이름을 지정하고 입술 영상을 저장한다. 예를 들어 “재생”란 발음을 하고 생성되는 입술 영상을 저장하기 위해서 파일 이름은 1_v513Lip으로 지정하고, 17프레임으로 저장된다면 각 프레임의 파일 이름은 1_v513Lip001부터 v513Lip016까지 프레임이 저장되도록 하였다.

학습데이터는 위의 그림 17에서 저장된 입술 영상들에 PCA(Principle Component Analysis)를 적용하여 고유입술(EigenLips)을 구하고 한 프레임 당 특징 계수를 추출한다. 그림 18은 한 단어를 발성하고

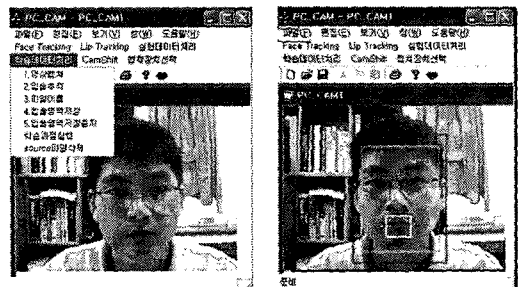


그림 16. 영상획득 및 얼굴과 입술 영역 검출

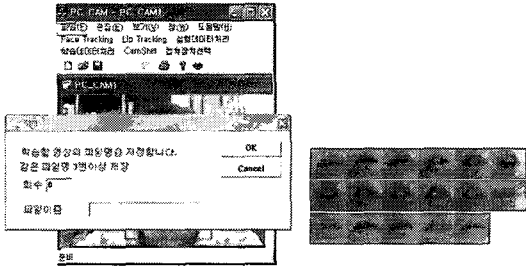


그림 17. 입술 영역 이미지 파일 저장

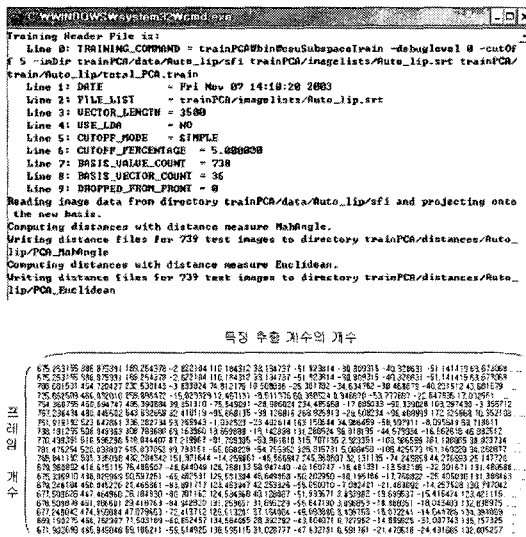


그림 18. 한 단어의 입술영상에 대한 주성분 분석 결과와 특징 계수 집합

생성되는 입술 영상들에 PCA를 적용하여 프레임별로 생성되는 특징 계수들을 구한 결과를 나타낸다. 학습 데이터로 사용된 입술 이미지의 크기가 70×50이므로 VECTOR_LENGTH는 3,500개가 되고, BASIC_VALUE_COUNT의 738은 프레임 0부터 738 프레임까지의 739개의 학습 데이터 전체 프레임의 개수를 나타낸다. 또한 BASIC_VECTOR_COUNT의 36은 각 프레임에서 추출되는 특징 계수의 수이다.

인식데이터의 처리는 실시간으로 입력되는 입술 영상의 각 프레임은 학습데이터에서 계산된 고유입술(EigenLips)과 연산을 하여 각 프레임의 특징 계수를 추출한다. 다음으로, 학습데이터와 인식데이터에서 구한 특징 계수를 HTK에 적용하여 학습과 인식을 수행한다. 그림 19는 “앞으로”라는 고립 단어를 발생했을 때 생성되는 입술 영상의 특징 계수를 이용

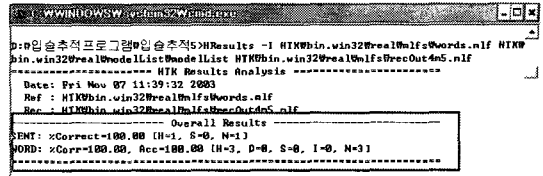


그림 19. 단어 인식데이터 HTK 실행 결과

하여 HTK에서 립리딩을 수행하고 인식 결과를 분석한 화면이다. 여기에서, SENTence 레벨인 경우 발생 단어의 개수가 1(Number=1), 정확히 인식된 단어의 개수가 1(Hit=1), 다른 단어로 대치되어 잘못된 인식된 단어의 개수가 0(Substitution error=0)임을 나타낸다. WORD 레벨인 경우 단어의 앞뒤에 존재하는 묵음까지 포함(sil 앞으로 sil)하므로 하나의 단어를 인식하기 위한 단어 개수는 3(Number=3), 정확히 인식한 단어의 개수가 3(Hit=3), 단어가 삭제되어 잘못된 인식된 경우가 0(Deletion error=0), 다른 단어로 대치되어 잘못된 인식된 단어의 개수가 0(Substitution error=0), 인식 목록에 없는 단어로 잘못 추가 인식된 단어의 개수가 0(Insertion error=0)임을 나타낸다.

이와 같이, 본 논문에서 구현한 립리딩 시스템은 영상획득에서부터 입술검출과 입술영역저장, 특징 계수 추출과 인식까지의 과정이 시스템 온라인 상태에서 한 번에 수행되도록 하였다. 실험에 사용된 학습데이터는 5개의 단어를 8회 발생하여 구현된 시스템에서 직접 영상을 획득했고, 인식데이터는 동일한 5개의 단어를 2회 발음한 영상을 사용했다. 실험 결과 화자종속 립리딩 시스템만의 인식률은 90%의 결과를 얻었다.

4. 결론 및 향후 연구

기존의 립리딩 실험에서는 입술 영상만을 이용한 제한된 실험데이터를 사용하거나 입술에 마커를 장착, 색칠을 하는 등, 인위적인 조작에 의하여 입술을 검출하였다. 또 배경과 조명을 고정하여 제한을 둔 입력영상으로 시스템 오프라인 상태에서 인식실험을 하였다. 본 논문의 립리딩 시스템은 기존의 립리딩 시스템의 제약성을 보완하고 하나의 독립된 시스템에서 립리딩 수행의 전체적인 과정이 시스템 온라인 상태에서 실시간으로 실행되도록 구성하였다. 실험 내용으로는 입력 영상에 제한을 두지 않은 영상으

로 입술 실험데이터를 추출하였고 시스템 오프라인 상태에서 실험되었던 립리딩 시스템을 시스템 온라인 상태에서 실시간으로 작동하도록 하여 사용자가 립리딩 과정의 학습과 인식을 직접 할 수 있게 하였다. 립리딩 시스템 과정에서 얼굴 영역 검출은 색도 히스토그램 모델을 이용하여 얼굴 영역 확률 분포를 구하였고 얼굴 영역 검출과 추적을 위해서는 평균 이동 알고리즘을 이용하였다. 입술 영역 검출에는 색도와 명도 값을 이용하였다. 검출된 입술 영역에서 PCA를 사용하여 입술 영역의 특징 계수를 구하였고 계산된 특징 계수를 HTK 형식의 파일 포맷으로 변환하여 HTK에 적용, 학습 및 인식을 하였다. 제안한 립리딩 시스템을 이용하여 실험을 수행 한 결과, 잡음이 있는 음성인식 환경에서의 인식률 저하를 보상할 수 있었으며, 입술 정보만 이용한 화자종속 립리딩 시스템만으로 90%의 인식률을 보였다. 또한 저가의 범용 PC 카메라를 사용하여 저조한 화질에서도 립리딩 시스템이 가능함을 보였다. 하지만 각 입술영상에서 추출되는 특징 계수의 계산 처리시간이 적지 않게 요구되는 단점도 있었다. 향후 좀 더 적은 양의 신뢰성 있는 특징 계수 추출방법이 지속적으로 연구된다면 더욱 효과적으로 실시간 처리가 가능한 립리딩 시스템이 구현될 것으로 생각된다.

참 고 문 헌

[1] C. Bregler, Y. Konig, "Eigenlips for Robust Speech Recognition", Processings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 669-672, 1994.

[2] P. L. Silsbee, A. C. Bovik, "Computer Lip-reading for improved accuracy in automatic speech recognition", IEEE Trans. Speech Audio Processing, vol. 4, no, 5, pp. 337-351, 1996.

[3] G. I. Chiou, J. N. Hwang, "Lipreading from Color Video", IEEE Transactions on Image Processing, Vol. 6, No. 8, pp. 1192-1195, Aug. 1997.

[4] U. Meier, R. Stiefelwagen, J. Yang, A. Waibel, "Towards Unrestricted Lipreading", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 14, No. 5, pp. 571-

785, 2000.

[5] C. Neti, G. Potamianos, J. Luetin, I. Matthews, H. Glotin, D. Vergyri, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins Summer 2000 Workshop", Multimedia Signal Processing, 2001 IEEE Fourth Workshop on, pp. 619-624, 2001.

[6] Gary R. Bradski, "Real Time Face and Object Tracking as a Component of a Perceptual User Interface", Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on, 19-21, pp. 214-219, Oct. 1998.

[7] W.T. Freeman, K. Tanaka, J. Ohta, K. Kyuma, "Computer vision for computer games", Automatic Face and Gesture Recognition, 1996, Proceedings of the Second International Conference on, 14-16, pp.100-105, Oct. 1996.

[8] Tsuhan Chen, R.R. Rao, "Audio-visual integration in multimodal communication", Proceedings of the IEEE, Volume: 86 Issue: 5, pp.837-852, 1998.

[9] H. McGurk, J. MacDonald, "Hearing lips and seeing voices", Nature, vol. 264, pp. 746-748, 1976.

[10] J.S.D. Mason, J. Brand, R. Auckenthaler, F. Deravi, C. Chibelushi, "Lip Signatures for automatic person recognition", multimedia Signal Processing, 1999. IEEE 3rd Workshop on, pp.457-462, 1999.

[11] M. Lievin, F. Luthon, "Lip features automatic extraction", Image Processing, ICIP 98. Proceedings. 1998 International Conference On, Vol.3, pp.168-172, 1998.

[12] 이은숙, 이호근, 이지근, 김봉완, 이상설, 이용주, 정성태, "건고한 입술 영역 추출을 이용한 립리딩 시스템 설계 및 구현", 한국 멀티미디어 학회 춘계 학술발표대회 논문집, pp.524-527, 2003.

[13] 이은숙, 이지근, 이상설, 정성태, "PCA와 HMM을 이용한 화자종속 립리딩 시스템", 한국 정보과학회 호남·제주지부 학술발표대회 논문집, pp.139-142, 2003.



이 지 근

2000년 2월 원광대학교 컴퓨터 공학과 졸업
2002년 2월 원광대학교 컴퓨터 공학과 석사학위 취득
2002년~현재 원광대학교 컴퓨터공학과 박사과정

관심분야: 영상인식, 영상처리, 컴퓨터 그래픽스, HCI



정 성 태

1987년 2월 서울대학교 컴퓨터 공학과 졸업
1989년 2월 서울대학교 컴퓨터 공학과 석사학위 취득
1994년 8월 서울대학교 컴퓨터 공학과 박사학위 취득
1995년 3월~현재 원광대학교 전기전자및정보공학부 교수

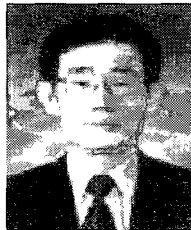
관심분야: 영상인식, 영상처리, 컴퓨터 그래픽스, HCI, VLSI



이 은 속

2002년 2월 원광대학교 컴퓨터 및정보통신공학부 졸업
2004년 2월 원광대학교 컴퓨터 공학과 석사학위 취득
2004년~현재 (주) 뮤콤 개발팀 연구원

관심분야: 영상처리, 영상인식, DMB



이 상 설

1984년 2월 고려대학교 전자공학과 졸업
1989년 2월 한국과학기술원 전기및전자공학과 석사학위 취득
1994년 2월 한국과학기술원 전기및전자공학과 박사학위

취득
1994년~현재 원광대학교 전기전자및정보공학부 교수
관심분야: 병렬컴퓨터구조, SoC, 영상및통신VLSI, 임베디드시스템