(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년06월03일
(11) 등록번호 10-2405570
(24) 등록일자 2022년05월31일

(51) 국제특허분류(Int. Cl.)
G10L 15/25 (2013.01) *G06K 9/00* (2022.01)
G06V 10/40 (2022.01) *G10L 15/08* (2006.01)
(52) CPC특허분류
G10L 15/25 (2013.01)
G06V 10/40 (2022.01)
(21) 출원번호 10-2020-0015368
(22) 출원일자 2020년02월10일
심사청구일자 2020년02월10일
(65) 공개번호 10-2021-0101413
(43) 공개일자 2021년08월19일
(56) 선행기술조사문헌
JP2011191423 A*
KR1020160059768 A*
KR1020160064565 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
대구대학교 산학협력단
경상북도 경산시 진량읍 대구대로 201 (대구대학교)
(72) 발명자
김성우
대구광역시 동구 동호로2길 3-3, 505호
차경애
대구광역시 수성구 달구벌대로 3304-1, 샤월화성파크드림
박세현
대구광역시 수성구 청호로84안길 2, 103동 1006호
(74) 대리인
이선택

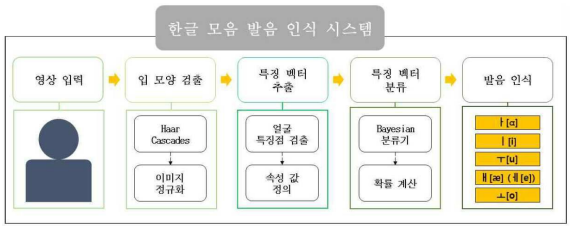
전체 청구항 수 : 총 2 항 심사관 : 임민섭

(54) 발명의 명칭 **베이지안 분류를 이용한 입 모양 기반의 발음 인식방법**

(57) 요 약

입 모양 기반의 발음 인식방법은, 입력영상에서 입 모양을 검출하고, 이미지 크기를 정규화 하여 얼굴 크기와 상관없이 일정한 특징점을 표현하는 속성 값을 정의하는 단계와, 베이지안 분류기를 통해 특징점을 표현하는 속성 값에 따른 한글 모음 발음 인식단계를 포함하는 것을 특징으로 한다.

대 표 도 - 도1



(52) CPC특허분류
G06V 40/16 (2022.01)
G10L 15/08 (2013.01)

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

삭제

청구항 2

삭제

청구항 3

삭제

청구항 4

입력영상에서 입 모양을 검출하고, 이미지 크기를 정규화 하여 얼굴 크기와 상관없이 일정한 특징점을 표현하는 속성 값을 정의하는 단계; 및

베이지안 분류기를 통해 특징점을 표현하는 속성 값에 따른 한글 모음 발음 인식단계;를 포함하고,

속성 값을 정의하는 단계는, 상기 입력영상에서 Haar-Cascades 알고리즘을 사용하여 검출된 얼굴에 특징점을 부여하고 입 모양을 검출함에 있어서, 상기 입력영상에서 추출한 이미지 크기를 정규화 하여 입 모양 특징점으로 정의한 요소들 사이의 거리를 계산하여 베이지안 학습 모델에 적용시킬 속성 값을 정의하고,

'ㅏ[a]', 'ㅣ[i]', 'ㅜ[u]', 'ㅐ[æ](ㅔ[e])', 'ㅗ[o]' 다섯 가지 한글모음을 구분할 수 있는 상기 입 모양 특징점을 정의하고,

상기 입 모양 특징점을 표현하는 속성 값은, 입 영역의 가로 길이[M_x], 입 영역의 세로 길이[M_y], 윗입술의 길이[L_a], 아랫입술의 길이[L_b], 입이 움직일 때 함께 변화를 보이는 턱과 볼을 사이의 거리를 이용하여 왼쪽 볼과 오른쪽 볼 사이의 거리[F_x], 입 영역의 상단과 턱의 하단 사이의 거리[F_y]로 정의되고,

상기 한글 모음 발음 인식단계는, 수학식 (2)와 같이 한글 모음 다섯 개를 각각 사건 클래스 C_i로 정의하고, 입 모양 특징점으로 검출된 속성 값(M_x,M_y,L_a,L_b,F_x,F_y)을 하나의 사건인 랜덤 변수 x로 정의하고,

<수학식 2>

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)}, i = 0, 1, ..., 4 \qquad (2)$$

$$C_0 = [a], C_1 = [i], C_2 = [u], C_3 = [æ], C_4 = [o]$$

$$X = (M_x, M_y, L_a, L_b, F_x, F_y)$$

속성값[M_x]을 크기에 따라 'ㅜ[u]' 발음과 'ㅏ[a]'발음, 그리고 'ㅐ[æ](ㅔ[e])' 발음을 구별하는 것을 특징으로 하는 입 모양 기반의 발음 인식방법.

청구항 5

삭제

청구항 6

제4항에 있어서,

ROI(Region Of Interest)를 이용한 이미지 정규화 방법;

속성 값들의 비율에 따른 이미지 정규화 방법;

추출된 얼굴 크기와 입의 크기를 기준으로 다른 속성 값들을 비교하는 이미지 정규화 방법; 및

한 개의 이미지를 기준으로 다른 모든 이미지들을 비교하는 이미지 정규화 방법; 중 어느 하나의 방법을 통해 이미지 크기를 정규화 하는 것을 특징으로 하는 입 모양 기반의 발음 인식방법.

청구항 7

삭제

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 발음인식방법에 관한 것으로서, 영상 정보를 활용한 얼굴 검출 및 입 모양 검출 기술에 관한 것이다.

배 경 기 술

[0003] IT기술의 발달로 영상 정보를 활용한 얼굴 검출 또는 입 모양 검출, 그리고 음성 정보를 활용한 음성 인식 기술이 보편화 되고 관련 어플리케이션이 개발됨에 따라 사람의 입 모양을 이용한 발음 인식에 관한 다양한 연구가 진행되고 있다.

[0004] 입 모양을 이용한 발음 인식 연구는 대부분 영상 정보와 음성 정보를 결합한 발음 인식 기술이 대부분이며, 최근에는 CNN(Convolutional Neural Network)이나 SVM(Support Vector Machine)같은 딥 러닝 기술을 접목시켜 영상에서의 발음 인식에 적용하는 추세이다. 하지만 딥러닝 기술은 시스템의 고성능화와 대용량의 학습 데이터를 확보해야 하는 제한 사항이 있으며, 음성 정보를 활용한 시스템에서는 주변 잡음이나 추가적인 정보 분석으로 시간적, 공간적 활용도가 떨어진다는 문제점이 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) KR 10-2014-0037410 A

발명의 내용

해결하려는 과제

[0007] 본 발명은 상기와 같은 기술적 과제를 해결하기 위해 제안된 것으로, 음성 정보나 딥 러닝 기술을 사용하지 않고 영상 정보만을 사용하여 실시간 영상에서 발음을 구별할 수 있는 베이지안 분류를 사용한 입 모양 기반의 발음 인식방법을 제공한다.

과제의 해결 수단

[0009] 상기 문제점을 해결하기 위한 본 발명의 일 실시예에 따르면, 실시간 영상에서 사람의 얼굴을 검출하고, 한글 모음 발화시 특징이 되는 입술 영역을 특징 벡터로 정의하여, 베이지안 분류기를 통해 ‘ㅏ[a]’, ‘ㅣ[i]’, ‘ㅜ[u]’, ‘ㅐ[æ](ㅔ[e])’, 'ㅗ[o]' 5가지 한글 모음 발음을 각각 분류하는 입 모양 기반의 발음 인식방법이 제공된다.

[0010] 또한, 본 발명에서 Haar-Cascades 알고리즘을 사용하여 사람의 얼굴을 검출하는 것을 특징으로 한다.

[0011] 또한, 본 발명에서 사람이 발화할 때 변화량이 큰 입과 입술, 그리고 볼 부위를 특징 벡터로 정의하여 베이지안 분류기를 통해 발음을 인식하는 것을 특징으로 한다.

[0013] 본 발명의 다른 실시예에 따르면, 입력영상에서 입 모양을 검출하고, 이미지 크기를 정규화 하여 얼굴 크기와 상관없이 일정한 특징점을 표현하는 속성 값을 정의하는 단계와, 베이지안 분류기를 통해 특징점을 표현하는 속성 값에 따른 한글 모음 발음 인식단계를 포함하는 입 모양 기반의 발음 인식방법이 제공된다.

[0014] 또한, 본 발명에서 Haar-Cascades 알고리즘을 사용하여 사람의 얼굴과 입 모양을 검출하는 것을 특징으로 한다.

[0015] 또한, 본 발명에서 ROI(Region Of Interest)를 이용한 이미지 정규화 방법, 속성 값들의 비율에 따른 이미지 정규화 방법, 추출된 얼굴 크기와 입의 크기를 기준으로 다른 속성 값들을 비교하는 이미지 정규화 방법, 한 개의 이미지를 기준으로 다른 모든 이미지들을 비교하는 이미지 정규화 방법 중 어느 하나의 방법을 통해 이미지 크기를 정규화 하는 것을 특징으로 한다.

[0016] 또한, 본 발명에서 속성 값을 정의하는 단계는, 상기 입력영상에서 Haar-Cascades 알고리즘을 사용하여 검출된 얼굴에 특징점을 부여하고 입 모양을 검출함에 있어서, 상기 입력영상에서 추출한 이미지 크기를 정규화 하여 입 모양 특징점으로 정의한 요소들 사이의 거리를 계산하여 베이지안 학습 모델에 적용시킬 속성 값을 정의하는 것을 특징으로 한다.

[0017] 또한, 본 발명에서 ‘ㅏ[a]’, ‘ㅣ[i]’, ‘ㅜ[u]’, ‘ㅐ[æ](ㅔ[e])’, 'ㅗ[o]' 다섯 가지 모음을 구분할 수 있는 입 모양 특징점을 정의하는 것을 특징으로 한다.

[0018] 또한, 본 발명에서 입 모양 특징점을 표현하는 속성 값은, 입 영역의 가로 길이[M_x], 입 영역의 세로 길이[M_y], 윗입술의 길이[L_a], 아랫입술의 길이[L_b], 입이 움직일 때 함께 변화를 보이는 턱과 볼을 사이의 거리를 이용하여 왼쪽 볼과 오른쪽 볼 사이의 거리[F_x], 입 영역의 상단과 턱의 하단 사이의 거리[F_y]로 정의되는 것을 특징으로 한다.

[0019] 또한, 본 발명에서 상기 한글 모음 발음 인식단계는, 수학식 (2)와 같이 한글 모음 다섯 개를 각각 사건 클래스 C_i로 정의하고, 입 모양 특징점으로 검출된 속성 값(M_x,M_y,L_a,L_b,F_x,F_y)을 하나의 사건인 랜덤 변수 x로 정의하는 것을 특징으로 한다.

[0020] <수학식 2>

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)}, i = 0, 1, ..., 4 \qquad (2)$$

$$C_0 = [a], C_1 = [i], C_2 = [u], C_3 = [æ], C_4 = [o]$$

$$X = (M_x, M_y, L_a, L_b, F_x, F_y)$$

[0021]

[0023] 또한, 본 발명에서 상기 한글 모음 발음 인식단계는, 데이터 집합은 정규 분포를 따른다고 가정하면, 랜덤 변수 X가 속할 확률분포인 한글 모음 중 하나를 결정하는 클래스 C_i 의 확률 밀도 함수의 평균이 u_i 공분산이 Σi 라고 할 때, 우도는 수학식 (3)으로 정의되는 것을 특징으로 한다.

[0024] <수학식 3>

$$P(X|C_i) = N(\mu_i, \sum_i) = \frac{1}{(2\pi)^{d/2}\left|\sum_i\right|^{1/2}} exp(-\frac{1}{2}(x-\mu_i)^T \sum^{-1}(x-\mu_i)) \qquad (3)$$

[0025]

[0026] 또한, 본 발명에서 상기 한글 모음 발음 인식단계는, 랜덤 변수 X가 입력되면 C_i 에 대해 사후 확률이 가장 큰 값으로 나타나는 클래스가 입 모양 특징점에 의해 분류되는 모음이고, 여기에서 수학식 (1)과 같이 사후 확률은 사전 확률과 우도의 곱으로 계산되므로, 최종적으로 우도가 최대인 클래스를 찾아 해당 모음 발음으로 분류하는 것을 특징으로 한다.

[0027] <수학식 1>

$$P(C_i|X) = \frac{P(C_i)P(X|C_i)}{P(X)}, i = 0, 1, ..., n \qquad (1)$$

[0028]

[0029] - P(C)와 P(X)는 각 사건 C와 사건 X가 일어날 사전 확률이며, P(X|C)는 사건 C가 일어났을 때, 사건 X가 발생할 확률을 의미하며 이를 우도(Likelihood)라 정의함 -

발명의 효과

[0031] 본 발명의 실시예에서는 영상 정보에서 입 모양의 변화를 반영하는 특징값을 이용하여 한글 모음 발음을 인식할 수 있는 베이지안 분류 기반의 알고리즘을 제안하였다.

[0032] ‘ㅏ[a]’, ‘ㅣ[i]’, ‘ㅜ[u]’, ‘ㅐ[æ](ㅔ[e])’, 'ㅗ[o]' 다섯 가지의 발음을 각각 하나의 클래스로 지정하고 베이지안 분류기에 적용시킨 입 모양 기반의 발음 인식방법을 적용하여 영상만을 사용한 발음 인식 시스템을 구현하였다.

[0034] 기존의 발음 인식 시스템은 음성 정보를 결합하거나 입 모양 검출을 위해 CNN이나 SVM 등 복잡한 계산 절차를 거쳐야 한다는 단점이 있다. 그에 비해 제안한 방법 및 시스템은 입 모양의 특징점 검출을 위한 픽셀 기반의 이미지 처리 과정을 간소화하고, 딥러닝 기법에 비교하여 계산 복잡성이 낮아 학습이 시간이 오래 걸리지 않고 높은 사양의 하드웨어를 요구하지 않는다는 장점을 가진다.

[0036] 제안한 방법 및 시스템에서 얼굴 검출에 사용한 Haar-Cascades 알고리즘은 다양한 Feature를 가지는 Haar-like Feature를 사용하여 사람의 얼굴에서 밝기 차를 계산, 입력 영상에서 정확하고 빠르게 얼굴 검출이 가능하다.

[0037] 또한 단순히 입의 좌표를 사용하지 않고 발화 시 특징이 되는 6가지 특징 벡터를 정의함으로서 독립성에 근거한 나이브 베이지안 알고리즘에 적용하여 정확한 발음 분류가 가능함을 보였다. 정의한 특징 벡터는 정규화 된 사진에서 좌표 값들의 합과 차의 계산만으로 빠르고 간단하게 검출이 가능하고 효과적으로 발음을 구별하기 위한 고유한 특징이 된다.

[0038] 실험을 통해서 입 모양 특징 벡터의 확률 분포가 다섯 가지 한글 모음 발음을 구분할 수 있는 모수 분포로 갱신되어지며, 초기 학습 데이터의 양이 적더라도 모음 발음을 인식할 수 있음을 보였다. 결과적으로 학습 데이터를 많이 필요로 하는 CNN을 사용한 유사 시스템보다 좋은 성능을 보였으며, 입 모양에 따른 발음 인식 시스템에 적

합하다고 판단된다.

[0039] 실험 결과로 최대 93%의 인식률을 보이며, 이러한 시각 정보만을 활용한 발음 인식 연구는 여러 플랫폼에 적용이 가능하여, 소음이 심한 환경이나 청각적 불편함이 있는 사람들에게 효과적인 인터페이스를 제공할 수 있으며, 자동 자막 생성과 같은 연구에 도움이 될 수 있다.

도면의 간단한 설명

[0041] 도 1은 본 발명의 실시예에 따른 한글 모음 발음 인식방법의 개념도

도 2는 Haar-Cascades 알고리즘에 대한 개념도

도 3은 얼굴 랜드마크 추출의 예시도

도 4는 속성 값의 정의

도 5는 실험에 사용된 영상의 예를 나타낸 사진

도 6은 얼굴 영역 검출과 랜드마크 부여 과정에 대한 도면

도 7은 속성값 [M_x] 의 확률밀도함수 그래프

도 8은 테스트에 사용된 이미지 영상의 예시도

도 9는 화자 종속 데이터에서의 속성 값 확률 분포를 나타낸 도면

도 10은 훈련 이미지에 따른 속성 [Mx] 값의 확률을 나타낸 도면

도 11은 훈련 이미지 개수에 따른 [Fx]속성 값의 확률을 나타낸 도면

도 12는 입력 영상에서 추출된 입 모양 영상

도 13은 각 발음 별 오인식율 그래프

도 14는 밝은 조명에서 촬영한 실시간 영상에서의 발음 검출 결과를 나타낸 도면

도 15는 어두운 조명에서 촬영한 실시간 영상에서의 발음 검출 결과를 나타낸 도면

발명을 실시하기 위한 구체적인 내용

[0042] 이하, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 정도로 상세히 설명하기 위하여, 본 발명의 실시예를 첨부한 도면을 참조하여 설명하기로 한다.

[0044] 본 발명에서는 Haar-Cascades알고리즘을 사용하여 얼굴을 빠르고 정확하게 검출하고, 입과 입술, 그리고 볼 등 사람이 발화할 때 변화량이 큰 얼굴 부위를 특징 벡터로 정의하여 베이지안분류기를 통해 발음을 인식하는 시스템 및 방법을 제안하였다.

[0045] 나이브 베이지안 분류기는 독립성에 기반한 확률 분류기로, 사전 확률과 사후 확률을 통해 훈련시킨 후 새로운 입력 영상에 대해 분류한다. 따라서 발화시 모양이 크게 변하지 않고 특징이 분명한 한글 모음 발음은 특징 벡터를 추출하여 나이브 베이지안 분류기에 적용 가능하다.

[0046] 제안하는 시스템 및 방법은 정적인 이미지를 수집하여 나이브 베이지안 분류기를 통해 훈련시킨 후, 정적인 영상과 실시간 영상에 적용시켜 음성 정보 없이 5개의 한글 모음인 ‘ㅏ[a]’, ‘ㅣ[i]’, ‘ㅜ[u]’, ‘ㅐ[æ](ㅔ[e])’, 'ㅗ[o]' 5개의 한글 모음 클래스를 분류하고 최고 90%의 인식률을 보여주었다.

[0048] 즉, 영상 정보를 이용한 발음의 인식을 위해서는 입 모양의 형태 변화를 특징값들의 상관관계를 분석하는 방식으로 가능하다. 이를 위해서 얼굴에서 특징이 뚜렷한 입 영역을 검출하여 특징 벡터로 정의하고 베이지안 분류기에 적용시켜 한글 모음 발음을 인식하는 시스템 및 방법에 대해서 설명한다.

[0050] 도 1은 본 발명의 실시예에 따른 한글 모음 발음 인식방법의 개념도이다.

[0051] 본 발명에서는 실시간 영상에서 사람의 얼굴을 검출하고 한글 모음 발화 시 특징이 될 수 있는 입술 영역을 특징 벡터로 정의하여 베이지안 분류기를 통해 ‘ㅏ[a]’, ‘ㅣ[i]’, ‘ㅜ[u]’, ‘ㅐ[æ](ㅔ[e])’, 'ㅗ[o]' 5가지 한글 모음 발음을 분류하는 방법을 제안한다.

[0052] 한글 모음 발음 인식 시스템 및 방법은 아래 도 1과 같이 영상입력, Haar-Cascades 알고리즘을 사용한 입 모양 검출, 특징 벡터 추출, 그리고 베이지안 분류기를 통한 특징 벡터에 따른 한글 모음 발음 인식으로 크게 총 4가지 과정으로 구성되어 있다.

[0055] **- Haar-Cascades classifiers를 사용한 얼굴 검출**

[0056] 도 2는 Haar-Cascades 알고리즘에 대한 개념도이다. - a) 4개의 Haar-like Feature, (b) Haar-like Feature로 영상을 탐색하는 모습.-

[0057] 본 발명에서 사용하고 있는 첫 번째 알고리즘은 얼굴 검출 알고리즘인 Haar-Cascades classifiers 알고리즘이다.

[0058] Haar-Cascades는 객체에서 보여주는 패턴을 분석하여 해당 객체를 찾는 알고리즘으로, 사람 얼굴 영역을 Haar-like Feature를 사용하여 검출한다. Haar-like feature는 Viola, Jones가 개발한 알고리즘이다. 영상에서 해당 영역과 다른 영역간의 밝기 차이를 이용하여 해당 객체를 검출하는 방법으로, 다양한 Feature형태가 존재한다.

[0059] 도 2와 같이 설정된 여러 가지 Feature로 Sliding Window 방식으로 영상을 탐색하여 얻은 Feature의 흰 부분의 밝기 합과 검은 부분의 밝기 합의 차이로 특징 값을 얻을 수 있다. 이렇게 Haar-Cascades 알고리즘은 복잡한 연산이 필요하지 않은 단순 합과 차의 연산을 사용하여 얼굴 검출 알고리즘에서 빠르고 정확하게 얼굴을 검출할 수 있다는 장점이 있다.

[0061] Cascades는 여러 개의 검출기를 순차적으로 사용하는 기법으로 처음에는 간단한 검출기를 적용한 뒤, 이후 더 복잡한 검출기를 적용해 나가는 방식이다. Cascades방식은 쉽게 검출 할 수 있는 특징들을 간단한 검출기로 빠르게 검출시켜 영상에서 객체를 검출하는 속도를 향상시킬 수 있다는 장점이 있다.

[0063] **- 입 모양 특징점 검출**

[0064] 도 3은 얼굴 랜드마크 추출의 예시도이다. - (a) Haar 알고리즘에 의해 검출된 얼굴 객체들, (b) 추출된 랜드마크-

[0065] 본 발명에서는 영상에서 얼굴을 감지하고 입 영역을 추출하기 위해서 Haar 알고리즘을 사용하였다. 또한 dlib에 정의되어있는 학습 모델을 이용하여 검출된 얼굴에 특징점을 부여한다. 도 3의 (a)는 Haar Cascades방식으로 인식된 얼굴과 검출된 눈, 입을 표시한 것이다. 검출된 영역에서 도 3의 (b)와 같이 dlib의 학습 모델을 이용하여 총 68개의 랜드마크를 추출한다.

[0066] 특징점을 찾은 후 이미지 크기를 정규화 하여 입 모양 특징점으로 정의한 요소들 사이의 거리를 계산하여 베이지안 학습 모델에 적용시킬 속성 값을 정의한다. 입력 영상에서 추출된 이미지를 정규화 과정을 거쳐 영상에서의 얼굴 크기와 상관없이 일정한 특징점을 표현하는 속성 값을 구할 수 있다.

[0068] 한글 발음에 있어서 입 모양 패턴은 대부분 모음에 의존하며 그 중에서도 기본이 되는 것은 단모음으로 입술의 움직임 정도에 따라서 발음이 구분된다. 본 발명에서는 한글의 기본형 모음 중 'ㅏ[a]', 'ㅣ[i]', 'ㅜ[u]', 'ㅗ[o]'의 네 개의 모음과 영어 모음에서도 사용되는 'ㅐ[æ](ㅔ[e])'의 총 다섯 가지의 모음을 구분할 수 있는 입 모양 특징점을 정의하고자 한다. 모음 발음이 시각 정보로 구분될 수 있는 것은 자음과 결합한 소리까지 인식하는데 있어서 매우 중요한 요소로 작용하기 때문이다.

[0070] 도 4는 속성 값의 정의를 나타낸 도면이다.

[0071] 한글 모음 발음 시 입 모양은 위 입술과 아래 입술의 좌우, 위아래 움직임의 변화가 있으며, 둥근 모양의 정도가 달라지면서 발음이 구분된다. 본 발명에서 정의한 입 모양 특징점은 도 4와 같이 입 영역의 가로 길이[M_x], 입 영역의 세로 길이[M_y], 윗입술의 길이[L_a], 아랫입술의 길이[L_b]로 정의한다.

[0072] 또한 인식률을 높이기 위해서, 입이 움직일 때 함께 변화를 보이는 턱과 볼을 사이의 거리를 이용하여 왼쪽 볼과 오른쪽 볼 사이의 거리[F_x], 입 영역의 상단과 턱의 하단 사이의 거리[F_y]를 측정하여 여섯 가지의 속성 값을 사용한다.

[0074] 이와 같은 입 모양의 속성 값들은 발음마다 높은 확률로 나타나는 해당 특징점의 근사 값을 구할 수 있어, 모음 발음을 판별하는 척도가 될 수 있다. 예를 들어, 입이 수평 방향으로 벌어지는 'ㅣ[i]'와 'ㅐ[æ](ㅔ[e])'의 경우에는 [M_x]와 [F_x]의 속성 값이 큰 폭으로 커지게 된다. 입이 수직 방향으로 벌어지는 'ㅏ[a]'와 'ㅗ

[o]' 의 경우는 [M_y]와 [F_y]의 속성 값이 커지게 된다.

[0075] 또한 [F_x]의 값을 통해서, 'ㅏ[a]' 에서 'ㅣ[i]' 나 'ㅐ[æ](ㅖ[e])' 발음으로 변할수록 양볼이 넓어지는 특성을 반영하여 'ㅏ[a]' 와 'ㅐ[æ](ㅖ[e])' 를 구별할 수 있다. 이는 베이지안 분류기의 학습 결과를 통해서 모음 발음을 구분할 수 있는 효과적인 특징점으로 사용됨을 볼 수 있다.

[0077] **- 베이지안 분류 기반 한글 모음 인식**

[0078] 한글 모음의 발음 인식을 위해서 입 모양을 구분하기 위한 알고리즘으로 베이지안 분류 기법을 사용한다. 이를 통해서 입 모양 특징점으로 검출되는 속성 값들의 사전 확률(Prior Probability)과 사후 확률(Posterior Probability) 분포에 근거하여 훈련과 테스트 과정이 반복되면서 분류하고자 하는 모음의 확률 분포가 갱신되어 가는 모델로 정의한다. 이는 적은 수의 학습 데이터로도 다섯 가지의 한글 모음 구분이 가능하고 화자 독립이나 화자 종속의 경우에 모두 적용할 수 있는 학습 모델이다.

[0080] 베이즈 정리는 수학식 (1)과 같으며, 여기서 P(C|X)는 사건 X가 일어난다고 가정했을 때 C가 일어날 조건부 확률로 이를 사후 확률이라고 한다.

[0081] P(C)와 P(X)는 각 사건 C와 사건 X가 일어날 사전 확률이며, P(X|C)는 사건 C가 일어났을 때, 사건 X가 발생할 확률을 의미하며 이를 우도(Likelihood)라고 한다

[0083] <수학식 1>

[0084]
$$P(C_i|X)=\frac{P(C_i)P(X|C_i)}{P(X)},i=0,1,...,n \qquad (1)$$

[0086] 본 발명에서는 수학식 (2)와 같이 한글 모음 다섯 개를 각각 사건 클래스 C_i로 정의하고, 입 모양 특징점으로 검출된 속성 값(M_x,M_y,L_a,L_b,F_x,F_y)을 하나의 사건인 랜덤 변수 x로 정의한다.

[0087] <수학식 2>

[0088]
$$P(C_i|X)=\frac{P(C_i)P(X|C_i)}{P(X)},i=0,1,...,4 \qquad (2)$$

$$C_0=[a],C_1=[i],C_2=[u],C_3=[æ],C_4=[o]$$

$$X=(M_x,M_y,L_a,L_b,F_x,F_y)$$

[0090] 데이터 집합은 정규 분포를 따른다고 가정하면, 랜덤 변수 X가 속할 확률분포인 한글 모음 중 하나를 결정하는 클래스 C_i 의 확률 밀도 함수의 평균이 u_i 공분산이 Σi 라고 할 때, 우도는 수학식 (3)이 된다.

[0091] <수학식 3>

[0092]
$$P(X|C_i)=N(\mu_i,\sum_i)=\frac{1}{(2\pi)^{d/2}\left|\sum_i\right|^{1/2}}exp(-\frac{1}{2}(x-\mu_i)^T\sum^{-1}(x-\mu_i)) \qquad (3)$$

[0093] 랜덤 변수 X가 입력되면 C_i 에 대해 사후 확률이 가장 큰 값으로 나타나는 클래스가 입 모양 특징점에 의해 분류되는 모음이다. 여기에서 수학식 (1)과 같이 사후 확률은 사전 확률과 우도의 곱으로 계산되므로, 최종적으로 우도가 최대인 클래스를 찾아 해당 모음 발음으로 분류한다.

[0095] **- 실험 환경**

[0096] 도 5는 실험에 사용된 영상의 예를 나타낸 사진이다. - 예 : (a) 'ㅏ[a]' 발음, (b) 'ㅐ[æ](ㅖ[e])' 발음 -

[0097] 본 발명에서 실험에 사용되는 이미지는 사람 연령과 성별에 구애받지 않도록 구성하였다. 실험에 사용된 영상 데이터는 20대, 30대, 그리고 50대 후반의 남, 여 10명의 발성 모습을 녹화한 총 500개로 구성된다.

[0098] 500개의 데이터는 'ㅏ[a]', 'ㅣ[i]', 'ㅜ[u]', 'ㅐ[æ](ㅖ[e])', 'ㅗ[o]' 의 다섯 개 모음 별로 각각

100개 씩으로 이루어져 있다.

[0099] 학습 모델을 위한 훈련 이미지는 다소 일정한 조명 환경에서 촬영된 이미지를 사용하였으며 실험 데이터의 경우 밝은 실내와 어두운 실내에서 촬영한 이미지로 30초 이내의 발음 영상을 사용하였다. 단 어두운 실내의 경우 화자의 얼굴이 명확하게 보일 정도의 조명 상태를 유지하였다.

[0101] 제안하는 입 모양에 의한 한글 모음 인식 시스템의 검증을 위해 표 1과 같은 환경에서 실험하였다. 하드웨어 구성은 Intel Core i7-3770 3.40GHz CPU와 Geforce GTX1060 3GB 그래픽 카드, 그리고 입력 영상의 해상도는 1090x1080로 구성하였다.

[0103] <표 1>

시스템	사양
CPU	Intel Core i7-3770 3.40GHz
Graphic Card	Geforce GTX1060 3GB
RAM	16GB
OS	Windows 10 Pro
Tool	python 3.6.5
Library	OpenCV 2.4.10
Camera	IPhone XS Camera

[0104]

[0106] - 이미지 정규화

[0107] 이미지에서 추출한 속성 값들은 이미지에서 추출된 얼굴의 모양 혹은 데이터의 추출 방법에 따라 값들이 달라진다. 얼굴의 모양은 얼굴의 각도, 얼굴의 크기, 그리고 얼굴의 위치 등에 따라 같은 발음이나 사람이라도 달라질 수 있다.

[0108] 따라서 각 이미지를 정규화시키는 작업이 필요하다. 이미지를 정규화시키는 방법은 하나의 이미지를 기초로 하여 다른 이미지를 기초 이미지와 똑같은 조건으로 만드는 방법 등 다양한 방법이 존재한다.

[0109] 본 발명에서는 각 속성 값들을 정규화 시키기 위해 테스트 방법을 4개로 나누어 각 정규화 방법에 따라 훈련 데이터를 생성할 때까지 걸리는 시간과 발음 분류 결과를 측정하는 테스트를 진행하였다.

[0110] 모든 테스트에서 사용된 이미지는 훈련 데이터 450개와 테스트 데이터 50개의 이미지로, 다섯 가지의 발음이 같은 비율로 구성되어 있다. 각 테스트의 진행은 다음과 같다.

[0112] 1) ROI(Region Of Interest)를 이용한 이미지 정규화

[0113] 첫 번째는 ROI(Region Of Interest)를 사용한 방법이다. ROI는 관심영역이라는 뜻으로, 영상처리 작업에서 영상 전체가 아닌 특정 영역에 대해 사용자가 원하는 처리를 수행하는 방법이다. 본 실험에서는 영상에서 사람의 얼굴을 검출한 다음, 얼굴 부분의 사각형영역을 추출하여 ROI로 지정한 후, 500 x 700 크기의 사각형에 ROI를 복사하여 이미지의 크기를 정규화 하였다.

[0115] 2) 속성들의 비율에 따른 이미지 정규화

[0116] 두 번째는 속성 값들의 비율을 이용한 방법이다. 각 사람이 가진 고유의 입모양과 크기는 사람마다 차이가 있기 때문에 같은 발음이라도 화자에 따라 입이 벌어지는 크기가 다를 수 있다. 하지만 각 사람이 발음을 하기 위해 사용하는 입이나 얼굴모양은 발음에 따라 비슷한 모양을 가진다. 따라서 하나의 속성 값을 기준으로 다른 속성 값들과 비교하여 각 속성 값들의 비율 값을 계산한다. 계산된 비율 값을 통해 이미지에서 추출된 얼굴 모양에 관계없이 비교 가능한 새로운 속성 값들을 정의하였다.

[0118] 3) 화자 얼굴 면적에 따른 이미지 정규화

[0119] 세 번째는 추출된 얼굴 크기와 입의 크기를 기준으로 다른 속성 값들을 비교하는 방법이다. 사람은 얼굴크기와 입의 크기가 다양하고 얼굴 모양이나 입모양 또한 사람에 따라 가지각색이다. 얼굴이 가로로 긴 사람은 같은 ‘ㅏ[a]’ 발음에도 얼굴이 세로로 긴 사람보다 양 볼 사이의 거리가 클 것이다. 따라서 추출된 각 사람의 얼굴이

나 입의 크기를 기준으로 6개의 속성 값과 비율을 계산하여 새로운 6개의 속성 값으로 정의하였다.

[0121] 4) 기준 이미지에 따른 이미지 정규화

[0122] 마지막으로는 한 개의 이미지를 기준으로 다른 모든 이미지들을 비교하는 방법이다. 각 발음에서 첫 번째로 훈련에 사용된 사람의 얼굴의 속성 값을 추출하여 해당 발음에서의 다른 사람들의 속성 값들의 변화 값을 검사한다.

[0123] 예를 들어, 추출된 이미지가 'ㅏ[a]' 발음이고 비교하는 이미지가 'ㅣ[i]' 발음인 경우 턱과 입 사이의 거리는 줄어들게 되고 양 볼의 거리는 넓어지게 된다. 이와 같이 기준이 되었던 이미지의 속성 값과 다른 이미지들의 속성 값의 변화량에 따라 각 발음을 구별할 수 있는 기준이 된다.

[0125] 아래 표 2는 정규화 방법에 따른 학습 데이터 생성 시간과 발음 인식률, 그리고 정적인 이미지일 때의 발음 인식 실험 시간과 동적인 영상일 때의 발음 인식 실험 시간을 나타낸다.

[0126] <표 2>

실험 방법	학습 시간 (단위 : 초)	정적인 영상의 테스트 시간 (단위 : 초)	동적인 영상의 테스트 시간 (단위 : 초)	발음 인식률
1. ROI	30.14초	0.022	100.01	83%
2. 속성 값의 비율	21.02초	0.020	97.17	74%
3. 얼굴/입 크기	29.78초	0.021	99.18	63%
4. 하나의 이미지 기준	28.59초	0.022	98.58	62%

[0127]

[0128] 정규화 방법에 따른 결과로 'ROI'의 사용이 30.14초, 인식률 83%. 속성 값의 비율이 21.02초, 74%. 얼굴/입의 크기 기준 31.78초, 63%. 하나의 이미지를 기준으로 하였을 때 32.59초, 62%의 결과를 보였다.

[0129] 'ROI'의 사용은 다른 방법과 다르게 이미지를 잘라내어 해상도를 재조절해야 하는 단계가 추가되어 다른 방법들 보다 최대 9초가량 시간이 느리다는 단점이 있었다. 하지만 인식률이 가장 낮은 4번에 비해 21%정도의 증가율을 볼 수 있었다. 2번의 경우 가장 빠른 생성 시간과 두 번째로 가장 높은 74% 인식률을 보였다.

[0130] 3번과 4번의 경우 인식률이 낮고 생성 시간도 느린 결과가 나왔는데, 그 이유는 3번의 경우 이미지마다 얼굴과 입의 크기를 계산해야하기 때문에 시간이 느린 결과가 나타났고, 4번의 경우 한 사람의 이미지와 다른 사람들의 입의 위치, 크기, 모양 등의 차이점으로 인해 인식률이 낮게 나왔다고 볼 수 있다.

[0131] 본 실험을 통해 정규화 방법에 따라 학습 데이터 생성 시간은 21초부터 30초까지 다양하였지만 해당 학습 데이터를 400개의 정적인 이미지 인식 테스트에 사용한 실험시간은 1초 내, 동적인 30초짜리 영상에 사용한 실험 시간차이는 최대 3초로, 발음 인식률이 가장 높은 1번 방법인 'ROI'를 사용하는 정규화 방법이 가장 효과적인 방법이라는 것을 알 수 있다.

[0133] - 실험 방범 및 실험 결과

[0134] 본 발명에서는 영상에서의 사람의 발음 인식을 위해 한글 발음의 기준이 되는 'ㅏ[a]', 'ㅣ[i]', 'ㅜ[u]', 'ㅐ[æ](ㅔ[e])', 'ㅗ[o]' 5가지 모음을 클래스로 정의하여 얼굴에서 추출된 각 속성 값을 통해 5가지 한글 모음 발음을 구별하도록 하였다.

[0135] 이를 통해 유사 연구와 딥 러닝을 사용한 시스템과 비교하여 비슷하거나 더 높은 성능을 목표로 하여 실험을 진행하였다. 제안하는 발음 인식 시스템은 화자 종속 실험, 훈련 횟수 실험, 그리고 전체 데이터에 대한 인식 실험으로 총 3가지 방법으로 실행했으며, 각 실험에서는 동일하게 Haar-Cascades 알고리즘으로 얼굴 영역을 검출한 뒤, 얼굴에서 특징점을 추출하여 베이지안 분류기에 적용시키는 과정을 진행하였다.

[0137] 1) 입 모양 특징 벡터의 추출과 속성 값의 확률 분포

[0138] 본 발명의 실험에서는 OpenCV를 사용하여 사람의 얼굴 영역을 표시하고 원하는 속성 데이터를 추출하였다. OpenCV는 'Open Source Computer Vision'의 약자로, 실시간 컴퓨터 비전을 목적으로 만든 라이브러리의 한 종

류이다.

[0139] OpenCV를 통해 이미지의 이진화(Binariztion), 노이즈 제거, ROI 설정 등 이미지 처리에 필요한 알고리즘을 쉽게 처리할 수 있다는 장점이 있다.

[0140] Haar-Cascades 알고리즘은 Haar-like Feature를 통해 배경과 사람의 얼굴을 구분하고 사람의 얼굴에서 눈과 입을 찾는 알고리즘으로, Haar-Cascades 데이터가 미리 학습되어 있는 XML 파일을 제안하는 시스템에서 불러온 뒤 입력 영상에서 사람의 얼굴을 빠르고 정확하게 검출하도록 하였다.

[0141] 얼굴 영역 검출은 Haar-Cascades 학습 데이터와 OpenCV를 사용하여 gray 영상으로 변환된 이미지에서 detectMultiScale 함수를 통해 검출하였다.

[0143] 검출된 얼굴 영역은 각 4개의 좌표 점으로 변환되어 튜플에 (x, y, w, h) 형태로 저장된다. 여기서 x, y는 좌표점의 x좌표와 y좌표를 의미하며 w, h는 영역의 가로 길이와 세로 길이를 의미한다. 이후 검출된 얼굴 영역을 ROI영역으로 지정하고 500x700 사이즈의 이미지로 정규화 한다.

[0145] 도 6은 얼굴 영역 검출과 랜드마크 부여 과정을 나타낸 도면이다.

[0146] 다음은 검출된 얼굴에 대한 랜드마크 부여 및 속성 값 추출이다. dlib 라이브러리에 미리 훈련되어 있는 데이터를 불러와 정규화 된 이미지에 적용시킴으로서 검출된 얼굴에 랜드마크를 부여할 수 있다. 랜드마크 또한 (x, y)의 튜플 형식으로 저장되게 되며, 68개의 좌표 점을 리스트 변수에 저장되게 된다.

[0147] 랜드마크를 부여한 후에 정규화를 진행하게 되면, 저장되어있던 좌표 값과 실제 이미지의 점 위치가 달라지는 문제가 있기 때문에 정규화 작업이 끝난 이미지에 랜드마크를 부여하는 것이 중요하다

[0149] 이후 정의해놓은 속성 값에 맞는 얼굴 특징점들 사이의 거리를 검출된 얼굴 특징점 좌표로 계산하여 발음에 따른 튜플의 집합을 생성한다. 생성된 튜플의 집합은 베이지안 분류기에 속성 값으로 적용시켜 학습시키고, 새로운 데이터가 들어왔을 때 나이브 베이즈 기반의 확률 모델로서 해당되는 발음을 추정할 수 있다. 아래의 표 3에서 표 7까지는 발음 클래스별로 추출한 5개의 이미지에서의 속성 값과 클래스 값을 나타낸 것으로, 각 클래스별 100개의 이미지를 사용하여 총 500개의 이미지에서 추출한 결과이다.

[0150] 표 3을 보면 'ㅏ[a]' 발음은 다른 발음에 비해 입이 수직 방향으로 벌어지는 값인 [M_y]와 [F_y]의 크기가 크다는 것을 알 수 있다.

[0151] 'ㅣ[i]' 발음은 입술의 두께인 [L_a], [L_b]의 값이 작고 양 볼의 길이 값인 [F_x]가 크다는 특징이 있는데, 사람이 'ㅣ[i]' 발음을 할 때 는 입이 옆으로 벌어지며 양 볼이 넓어지기 때문이다.

[0152] 비슷한 입 모양인 'ㅜ[u]'와 'ㅗ[o]' 발음의 큰 차이점은 입과 턱 사이의 길이인 [F_y]와 양 볼 사이의 길이인 [F_x]이다. 'ㅜ[u]' 발음에서 'ㅗ[o]' 발음으로 변할 때 사람은 입을 벌리게 되므로 양 볼의 수평 길이인 [F_x]는 작아지게 되고 입과 턱까지의 수직 길이인 [F_y]는 훨씬 커지므로, [F_x]와 [F_y]는 'ㅜ[u]'와 'ㅗ[o]'를 구별할 수 있는 속성 값이 된다.

[0153] 그리고 'ㅐ[æ](ㅔ[e])' 발음은 'ㅏ[a]'와 'ㅣ[i]' 발음의 중간 값을 띄고 있어 오인식을 불러오는 계기가 된다. 또한 'ㅐ[æ](ㅔ[e])' 발음은 가장 검출이 어렵기 때문에 많은 학습 데이터를 필요로 한다.

[0155] <표 3>

[0156] 'ㅏ[a]' 발음 클래스에 대해 추출된 속성 값 (단위 : px)

	[M_x]	[M_y]	[L_a]	[L_b]	[F_x]	[F_y]
ㅏ[ɑ] (1)	89.98	94.79	24.14	71.51	122.7	172
ㅏ[ɑ] (2)	99.79	113.85	23.38	94.79	162.27	191.11
ㅏ[ɑ] (3)	75.01	87.68	26.2	82.05	138.77	173.31
ㅏ[ɑ] (4)	77.83	93.47	26.88	80.62	135.95	179.45
ㅏ[ɑ] (5)	79.31	100.1	27.61	96.21	159.61	188.44

[0157]

[0159] <표 4>

[0160] ‘ㅣ[i]’ 발음 클래스에 대해 추출된 속성 값 (단위 : px)

	$[M_x]$	$[M_y]$	$[L_a]$	$[L_b]$	$[F_x]$	$[F_y]$
ㅣ[i] (1)	119.05	101.83	24.08	43.23	140.15	201.48
ㅣ[i] (2)	116.67	107.62	23.38	44.64	140.12	202.38
ㅣ[i] (3)	116.67	98.41	21.22	52.35	135.06	202.47
ㅣ[i] (4)	120.51	106.07	24.21	44.12	147.05	203.13
ㅣ[i] (5)	118.8	103.23	20.52	56.59	152.82	202.41

[0161]

[0163] <표 5>

[0164] ‘ㅜ[u]’ 발음 클래스에 대해 추출된 속성 값 (단위 : px)

	$[M_x]$	$[M_y]$	$[L_a]$	$[L_b]$	$[F_x]$	$[F_y]$
ㅜ[u] (1)	74.9	96.66	24.08	50.99	121.05	197.57
ㅜ[u] (2)	76.58	102.65	25.46	48.09	120.21	197.44
ㅜ[u] (3)	75.07	107.76	28.32	56.61	135.95	198.09
ㅜ[u] (4)	75.67	100.82	25.47	52.33	127.99	201.55
ㅜ[u] (5)	75.87	98.6	25.47	52.33	127.3	195.35

[0165]

[0167] <표 6>

[0168] ‘ㅐ[æ](ㅔ[e])’ 발음 클래스에 대해 추출된 속성 값 (단위 : px).

	$[M_x]$	$[M_y]$	$[L_a]$	$[L_b]$	$[F_x]$	$[F_y]$
ㅐ[æ] (1) (ㅔ[e])	85.61	96.05	22.64	46.67	113.92	189.52
ㅐ[æ] (2) (ㅔ[e])	84.39	102.63	24.05	53.04	122.38	188.96
ㅐ[æ] (3) (ㅔ[e])	104.65	95.78	23.35	43.85	128.69	191.67
ㅐ[æ] (4) (ㅔ[e])	106.08	95.35	22.64	52.37	129.19	195.39
ㅐ[æ] (5) (ㅔ[e])	95.55	95.52	21.93	50.96	136.78	188.99

[0169]

[0170] <표 7>

[0171] ‘ㅗ[o]’ 발음 클래스에 대해 추출된 속성 값 (단위 : px).

	$[M_x]$	$[M_y]$	$[L_a]$	$[L_b]$	$[F_x]$	$[F_y]$
ㅗ[o] (1)	70.91	110.28	27.72	51.15	120.8	194.44
ㅗ[o] (2)	73.62	98.38	28.28	57.99	125.96	207.91
ㅗ[o] (3)	70.84	111.29	27.58	63.64	125.17	212.13
ㅗ[o] (4)	73.54	107.01	28.29	67.89	135.07	205.08
ㅗ[o] (5)	82.73	105.76	29.71	58.02	130.14	195.88

[0172]

[0174] 또한 본 발명에서는 각 한글 모음 클래스 속성 값들과 속성 값들의 관계가 가지는 의미를 시각적으로 나타내기 위하여 확률밀도함수를 사용하여 확률 분포도로 나타내었다. 베이지안 분류에서 데이터 집합은 정규 분포를 따

른다고 가정하였기 때문에 추출된 속성 값에서 확률밀도함수로 나타낼 수 있다.

[0175] 정규분포는 평균m과 표준편차σ로 주어지는 확률밀도함수를 가질 때, X는 정규분포를 따른다고 할 수 있으며 $X \sim N(m, \sigma^2)$로 나타날 수 있다. 정규분포의 식은 <수학식 4>와 같다.

[0176] <수학식 4>

$$f(X,m,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{(x-m)^2}{2\sigma^2}} \qquad (4)$$

[0177]

[0178] 위의 정규분포 식을 사용하여 각 클래스에 따른 모든 속성 값에 대한 정규분포 값을 계산하면 확률밀도함수 그래프의 형식을 나타내게 된다. 도 7은 6개 발음 클래스에 대한 [M_x]값의 확률밀도함수 그래프로서 입 모양 특징점들에 대한 실제 획득된 속성 값의 확률 분포를 나타낸다.

[0179] 여기서 입력 이미지는 500×700의 크기로 정규화 하였기 때문에 얼굴의 중앙 아래쪽에 위치하는 입영역에서 추출되는 픽셀 값들은 대략 70 픽셀부터 최대 120까지의 값을 나타내고 있다. 각 속성의 픽셀 값은 얼굴 랜드마크에서 입 영역으로 정의된 특징들의 픽셀로 측정되는 거리를 의미한다.

[0180] 확률밀도함수에서는 해당 속성 값으로 구별하기 용이한 발음 클래스에 대해시각적으로 파악할 수 있는데, 속성 [M_x] 값에서는 특히 'ㅜ[u]' 발음과 'ㅏ[a]'발음, 그리고 'ㅐ[æ](ㅔ[e])' 발음의 구별이 용이하다는 것을 알 수 있다.

[0181] [M_x] 속성값이 75 - 80의 값을 가질 때는 ㅜ발음, 90-95 의 값을 가질 때는 'ㅏ[a]' 발음, 그리고 103-108의 값을 가질때는 'ㅐ[æ](ㅔ[e])' 발음을 의미한다고 판단할 수 있다. 확률밀도함수에서는 빈도수의 수치가 높을수록 해당 되는 발음이 나올 확률이 높다는 것을 의미함으로 다른 발음 클래스 그래프와 겹치지 않거나 더 높은 수치를 나타낼수록 속성 값이 해당 발음을 의미한다는 것으로 해석할수 있다.

[0183] 또한 본 발명에서 제안한 시스템으로 두 가지의 실험을 추가로 실시하였다.

[0184] 첫 번째로는 화자 종속 실험에 의한 입 모양 특징 벡터의 확률 분포 실험으로, 화자 독립일 때와 화자 종속일 때의 발음 인식률의 차이에 대해 알아본다.

[0185] 두 번째는 훈련 데이터에 따른 확률 분포 실험으로, 훈련 데이터량에 따른 발음 인식률의 차이에 대해 알아본다. 두 가지 실험 모두 실험 방법에 의해 발음 인식률의 현저한 차이를 보였으며, 제안한 시스템에서 인식률이 최대가 되는 실험 조건에 대해 알 수 있었다.

[0187] 2) 화자 종속 실험에 의한 한글 모음 발음 인식 결과

[0188] 첫 번째는 동일인을 대상으로 하는 화자 종속인 경우와 동일하지 않은 10명의 인원을 대상으로 하는 화자 독립인 경우의 한글 모음 인식률 실험이다. 테스트에 사용된 훈련 이미지는 발음별 40장씩 총 240장의 이미지를 사용, 테스트 이미지는 발음별 5장씩 총 30장의 이미지를 사용하였다. 따라서 화자 종속실험에서는 한 사람의 이미지 240장으로 훈련한 후, 동일인의 30장의 이미지를 사용하여 테스트하였다. 도 8은 테스트에 사용된 영상의 예를 보여준다.

[0190] 아래의 표 8은 훈련 데이터를 사용하여 무작위로 10회 반복하여 테스트 한 후 획득한 인식률의 평균을 계산한 결과이다. 결과로 화자 종속의 경우 94%의 발음 인식률을 보였다. 같은 화자의 경우 고유의 발음 모양을 가지기 때문에 발음에 따른 속성 값의 차이가 분명하고 한 가지 발음에 대한 속성 값의 변화가 크지 않아 인식률이 매우 높다. 반대로 화자 독립의 경우 사람에 따라 얼굴 모양과 발음 모양이 다르기 때문에 화자 종속의 경우보다 낮은 74%의 발음 인식률을 보였다. 또한 화자 종속의 경우 발음 인식률이 오차범위 5%로 일정한 인식률을 보였지만, 화자 독립의 경우 훈련 데이터에 사용된 이미지에 따라 테스트 결과가 다르게 나타나는 양상을 보여 인식률의 결과가 최대 90% 부터 최소 65%까지 크게 변화한다는 특징이 있었다.

[0192] <표 8>

[0193] 1인 이미지와 다수 이미지로 실험하였을 때의 발음 인식률

[0194]

대상 이미지	화자 종속	화자 독립
한글 모음 인식률	94%	74%

[0196] 또한 실험에 의한 결과를 나타낸 확률밀도그래프인 도 9를 통해 각 발음에서 해당 모음이 나올 수 있는 픽셀 값의 분포와 확률을 알 수 있다. - (a) 속성 [M_x] (b) 속성 [M_y], (c) 속성 [L_a].-

[0197] ‘ㅏ[a]’의 경우 픽셀 값이 93 픽셀이 나올 확률이 높으며 ‘ㅜ[u]’ 의 경우 80 픽셀, ‘ㅗ[o]’ 발음의 경우 105 픽셀의 값이 나올 확률이 높다는 것을 나타낸다.

[0198] 도 9의 (a)는 입의 가로 길이인 [M_x] 값에 대한 훈련 이미지에 따른 확률 분포로써 각 발음별 분포도가 뚜렷이 구별된다.

[0199] 특히 93 픽셀 값을 기준으로 ‘ㅜ[u]’와 ‘ㅗ[o]’ 발음이, ‘ㅐ[æ](ㅔ[e])’와 ‘ㅣ[i]’ 발음이 확실히 구분되는 것을 볼 수 있다. 도 9에서 (b) 속성 [M_y]의 경우, 발음 별로 속성 값들이 근사 값이 많이 나타났지만 (c) 속성 [La]의 경우 ‘ㅐ[æ](ㅔ[e])’ 발음과 ‘ㅗ[o]’ 발음이 높은 확률로 분류되는 현상이 나타났다.

[0200] 이는 베이지안 분류기의 입력인 랜덤변수 X= (M_x,M_y,L_a,L_b,F_x,F_y) □의 각 속성 값들의 확률 값을 모두 곱하여 우도를 계산하기 때문에 발음을 구분하는데 더 높은 확률로 나타나는 속성에 의해서 발음 인식이 가능하다는 것을 보여준다.

[0202] 도 10은 입의 가로길이인 속성 [M_x] 값에 대한 훈련 이미지에 따른 확률 분포도를 나타낸다. <(a) 화자 종속, (b) 다수 인원의 이미지>

[0203] 한 사람 이미지만 사용한 경우 각 발음별 [M_x] 속성값이 나올 확률이 값에 따라 뚜렷하게 보였으며 특히 93px으로 기준으로 발음 ‘ㅜ[u]’ 발음과 ‘ㅗ[o]’ 발음이 ‘ㅐ[æ](ㅔ[e])’ 발음과 ‘ㅣ[i]’ 발음이 확실히 구분되는 것을 알 수 있었다.

[0204] 또한 ‘ㅜ[u]’ 발음과 ‘ㅗ[o]’ 발음의 경우 다수 인원으로 실험을 하였을 때보다 한 사람으로 실험을 하였을 때 확률 분포의 교집합이 적어 80px에서 높은 확률로 ‘ㅜ[u]’ 발음으로 인식되어 정확도가 상승했다. 이처럼 다수 인원의 이미지를 사용했을 때보다 한 사람의 이미지를 사용했을 경우, 확률밀도의 교집합이 적어지고 확률 분포가 뚜렷하게 나타나 발음 인식률이 올라간다는 결과를 보였다.

[0206] 3) 훈련 횟수에 따른 한글 모음 발음 인식 결과

[0207] 베이지안 분류기에서는 변수가 많아짐에 따라 확률이 낮아진다는 특징이 있고, 각 랜덤 변수의 값이 다른 변수와 분명한 차이점을 가질수록 분류 성능이 높아진다는 특징이 있다. 너무 많은 데이터로 훈련할 시 훈련 과정에서 정제되지 않은 데이터들이 정확한 분류를 방해하여 오히려 성능이 떨어질 수도 있다. 하지만 많은 훈련 데이터를 사용할수록 더욱 정확하고 세밀한 데이터를 얻을 수 있게 되고, 정제된 데이터를 훈련 데이터로 사용하여 많은 데이터를 훈련하는 것은 분류 성능을 높이는 수단이 된다.

[0208] 두 번째 실험인 본 실험에서는 훈련 데이터 개수에 따라 발음이 구별되는 확률을 구하였다. 훈련 데이터 개수는 발음 당 10개, 20개, 40개, 60개로 총 50개, 100개, 200개, 300개로 설정하였으며 해당 훈련 데이터를 사용하여 정해진 50개의 이미지 데이터를 분류하였다.

[0210] <표 9>

[0211] 훈련 이미지 수에 따른 한글 모음 발음 인식률.

[0212]

훈련 이미지 수 (단위 : 개)	50	100	200	300
인식률(단위: %)	49%	68%	78%	82%

[0214] 분류 결과로는 50개의 훈련 데이터로 테스트 한 결과 전체 49%확률로 분류가 이루어졌고, 100개 68%, 200개 78%, 마지막으로 300개 훈련 데이터의 경우 82%의 인식률을 보였다. 테스트 결과로 보았을 때, 각 발음별 비슷한 속성 값이 많고 같은 발음이라도 화자에 따라 각 속성 값들의 변화가 크기 때문에 훈련 수가 적을수록 분류

성능이 떨어지고 훈련 수가 많을수록 더 정확한 분류가 가능한 것을 볼 수 있었다.

[0215] 도 11은 각 훈련 이미지 개수에 따른 [F_x]속성 값에 대한 확률 분포도를 나타낸 그림이다. - (a) 훈련 이미지 50개의 경우, (b) 훈련 이미지 300개의 경우 -

[0216] 훈련 데이터가 적을수록 확률 분포가 겹치는 부분이 많이 보이는데, 그것은 곧 분류 성능이 좋지 않다는 것을 의미한다.

[0217] 'ㅏ[a]' 발음과 'ㅣ[i]', 그리고 'ㅐ[æ](ㅔ[e])' 발음이 확실하게 구분되어 있지만 낮은 확률값을 보여준다. 낮은 확률 값은 다른 발음 속성 값과 비교했을 경우 같은 값이라도 확률적으로 높은 확률의 속성 값을 선택하게 되어 인식률의 저하를 불러일으킬 수 있다. 훈련 데이터가 많아질수록 속성 값들의 정확한 분류가 이루어지고 해당 속성 값이 나올 확률이 높아진다. 예를 들어 'ㅏ[a]' 발음의 경우 178px 값이 나올 확률이 가장 높고, 178px보다 낮은 값들은 다른 발음보다 'ㅏ[a]' 발음에서 해당 값이 나올 확률이 높으므로 높은 확률로 'ㅏ[a]' 발음으로 분류될 수 있다.

[0218] 본 실험을 통해 표 9에 보이는 것과 같이 훈련 데이터가 많아질수록 발음 인식에서 속성 값의 확률 증가와 발음 인식률 향상을 보여준다는 사실을 알 수 있다.

[0220] 4) 제안한 방법과 CNN 기반 한글 모음 발음 인식 결과 비교.

[0221] 위에서 기술한 실험 방법에 의해서 수집된 훈련 이미지를 무작위로 사용하여 제안한 베이지안 분류기 기반의 한글 모음 발음 인식 실험 결과와 CNN기반의 인식 실험을 비교하였다.

[0222] 사용된 이미지는 발음 클래스 별 100개의 이미지로 총 500개의 이미지를 사용하였으며, 그 중 발음 클래스 별 80개 이미지 총 400개의 이미지를 훈련 데이터, 나머지 100개 이미지를 테스트 데이터로 구분하여 실험을 진행하였다.

[0223] 한글 모음 발음 인식 실험은 실험의 정확도를 향상시키고 dlib에서 얼굴 랜드마크를 100% 정확히 추출할 수 있도록 카메라를 정면으로 봤을 때 10도 내외의 일정 각도에서 얼굴을 인식하도록 하였다. 얼굴 영역은 머리카락의 경계가 되는 이마부터 턱까지 인식 영역 내에 진입하도록 하였으며, 이를 통해서 랜드마크 추출은 매 영상마다 정확한 값을 보인다고 할 수 있다.

[0225] 도 12는 입력 영상에서 추출된 입 모양 영상이다. - (a) 'ㅏ[a]' 발음, (b) 'ㅐ[æ](ㅔ[e])' 발음, (c) 'ㅗ[o]' 발음 -

[0226] 실험을 위해 수집한 총 500개의 이미지를 데이터로 실험한 결과, 전체적으로 약 85%의 발음 인식률을 나타내었으며 표 10에서 보이는 바와 같이 'ㅏ[a]' 발음이 93%로 가장 높은 인식률을 나타내었다. 'ㅣ[i]' 발음이 85%로 두 번째로 높은 값을 보였으며, 가장 인식률이 낮은 것은 'ㅗ[o]' 발음으로 나타났다.

[0228] <표 10>

[0229] 한글 모음 발음의 인식 결과.

[0230]

모음 / 인식 결과	ㅏ[ɑ]	ㅣ[i]	ㅜ[u]	ㅐ[æ] (ㅔ[e])	ㅗ[o]
ㅏ[ɑ]	93%	0	0	0	0
ㅣ[i]	0	85%	0	16%	4%
ㅜ[u]	0	0	73%	0	17%
ㅐ[æ] (ㅔ[e])	0	15%	0	80%	0
ㅗ[o]	7%	0	27%	4%	79%

[0231] 도 13은 각 발음 별 오인식율 그래프이다.

[0232] 'ㅗ[o]' 발음이 가장 인식률이 낮은 이유는 'ㅜ[u]' 와 'ㅗ[o]' 의 경우 속성 값의 확률 분포가 다소 비슷한 양상을 보이기 때문에 'ㅜ[u]' 를 'ㅗ[o]' 로 오인식되거나 'ㅗ[o]' 를 'ㅜ[u]' 로 오인식되는 경우가 있기 때문이다. 따라서 'ㅗ[o]' 발음과 'ㅜ[u]' 발음은 테스트 데이터에 따라 인식률이 크게 차이가 나는 것을 볼 수 있다.

[0234] 아래의 표 11은 시각적 이미지를 분석하는데 사용되는 딥러닝의 한 종류인 CNN(Convolutional Neural Network)를 사용하여 본 발명에서 제안하는 시스템과 인식률을 비교한 표이다. CNN은 심층 신경망의 한 종류인 인공신경망으로서, Convolution Layer라고 하는 계층으로 연결되어 하나 또는 여러 개의 Convolution Layer로 구성된 다층구조를 가지고 있다. 또한 CNN은 2차원 데이터의 학습에 적합하며, 데이터가 다중 Convolution Layer를 통과하며 훈련을 거듭하여 최종적인 분류 결과에 도달하게 된다.

[0236] 본 발명에서는 이미지 인식에 주로 사용되는 딥러닝의 한 종류인 CNN을 비교함으로서 영상 이미지만을 사용한 한글 모음 발음 인식 시스템에서의 베이지안 분류기의 성능과 본 시스템의 유용성에 대해 알아보고자 하였다.

[0238] 표 11의 결과 값은 랜드마크가 거의 100%의 정확도로 추출되는 환경에서 실험된 결과이며 본 발명의 시스템과 CNN과의 비교를 위하여 CNN에 사용된 데이터와 실험 환경은 본 발명과 같은 환경에서 진행되었다.

[0239] 결과적으로 본 발명에서 제안한 시스템에서는 'ㅏ[a]' 발음이 93%로 가장 높은 인식률을 보였으며 CNN의 경우 'ㅜ[u]' 발음이 가장 높은 인식률을 보였다. 'ㅜ[u]' 발음을 제외한 모든 발음이 CNN을 사용한 경우보다 최대 26% 정도 높은 인식률을 보였으며, 이를 통해 기존 연구의 결과보다 향상된 인식 결과를 보인다는 것을 알 수 있다.

[0240] 또한 CNN의 경우에는 최대 80%에서 최소 30%까지 가장 낮은 인식률을 보였는데, 이는 5000개에서 10000개까지의 데이터의 효율성이 가장 좋은 CNN시스템에서 적은 수의 훈련 데이터 사용으로 인한 인식률 저하라고 볼 수 있다. 이는 본 시스템에서 제안한 베이지안 분류기를 사용한 발음 인식이 적은 수의 훈련 데이터를 사용했을 경우에도 높은 인식률을 보인다는 것을 의미한다.

[0242] 입술 검출 시스템에 따라 한글 모음 발음에서 강점과 약점이 뚜렷한 결과를 보였으나, 인식률이 최소 30%까지 떨어지는 다른 시스템들과는 달리 제안된 시스템의 경우 최대 90%에서 최소 75%의 인식률로 다른 시스템보다 한글 모음 발음 인식에 특화되어 있다는 것을 알 수 있었다.

[0244] <표 11>

[0245] CNN과 제안된 베이지안 분류기의 비교

[0246]

모음	ㅏ[a]	ㅣ[i]	ㅜ[u]	ㅐ[æ] (ㅔ[e])	ㅗ[o]
Bayesian	93%	85%	75%	80%	79%
CNN	50%	50%	80%	40%	30%

[0247] 본 발명에서는 정적인 이미지뿐만 아니라 동적인 실시간 영상에서의 발음인식 실험도 진행하였다.

[0248] 도 14는 실시간 영상에서 프레임마다 각 발음을 추출한 결과이다.- (a)발음'ㅏ[a]', (b) 발음'ㅣ[i]', (c) 발음'ㅐ[æ]', (d) 발음 'ㅗ[o]'.-

[0249] 도 15는 어두운 조명에서 촬영한 실시간 영상에서의 발음 검출 결과이다. - (a)발음'ㅏ[a]', (b) 발음'ㅣ[i]', (c) 발음 'ㅜ[u]', (d) 발음 'ㅐ[æ]' -

[0250] 실시간 영상은 30초의 'ㅏ[a]', 'ㅣ[i]', 'ㅜ[u]', 'ㅐ[æ](ㅔ[e])', 'ㅗ[o]' 다섯 개의 모음 발음을 발화하는 영상이며, 실시간 표정 변화와 입 모양에 따라 해당 발음을 시스템에서 인식하여 현재 발화중인 발음을 화면에 나타낼 수 있다.

[0252] 실험에서 사용된 영상 데이터는 영상은 아이폰 카메라로 촬영한 MPEG-4 동영상을 사용하였고, 인식된 발음을 영상의 상단부에 출력하여 영상에 대한 정보를 아래에 나타내었다. 훈련 데이터로는 위의 실험에서 사용 된500개 이미지를 미리 훈련하여 실험에 사용하였다. 그 결과로 조명에 관계없이 입 모양에 맞는 발음을 정확히 화면에 나타나는 것을 볼 수 있다.

[0254] 본 발명의 실시예에서는 영상 정보에서 입 모양의 변화를 반영하는 특징값을 이용하여 한글 모음 발음을 인식할 수 있는 베이지안 분류 기반의 알고리즘을 제안하였다.

[0255] 'ㅏ[a]', 'ㅣ[i]', 'ㅜ[u]', 'ㅐ[æ](ㅔ[e])', 'ㅗ[o]' 다섯 가지의 발음을 각각 하나의 클래스로 지정하고 베이지안 분류기에 적용시킨 입 모양 기반의 발음 인식방법을 적용하여 영상만을 사용한 발음 인식 시스템을 구현하였다.

[0257] 기존의 발음 인식 시스템은 음성 정보를 결합하거나 입 모양 검출을 위해 CNN이나 SVM 등 복잡한 계산 절차를 거쳐야 한다는 단점이 있다. 그에 비해 제안한 방법 및 시스템은 입 모양의 특징점 검출을 위한 픽셀 기반의 이

미지 처리 과정을 간소화하고, 딥러닝 기법에 비교하여 계산 복잡성이 낮아 학습이 시간이 오래 걸리지 않고 높은 사양의 하드웨어를 요구하지 않는다는 장점을 가진다.

[0259] 제안한 방법 및 시스템에서 얼굴 검출에 사용한 Haar-Cascades 알고리즘은 다양한 Feature를 가지는 Haar-like Feature를 사용하여 사람의 얼굴에서 밝기 차를 계산, 입력 영상에서 정확하고 빠르게 얼굴 검출이 가능하다.

[0260] 또한 단순히 입의 좌표를 사용하지 않고 발화 시 특징이 되는 6가지 특징 벡터를 정의함으로서 독립성에 근거한 나이브 베이지안 알고리즘에 적용하여 정확한 발음 분류가 가능함을 보였다. 정의한 특징 벡터는 정규화 된 사진에서 좌표 값들의 합과 차의 계산만으로 빠르고 간단하게 검출이 가능하고 효과적으로 발음을 구별하기 위한 고유한 특징이 된다.

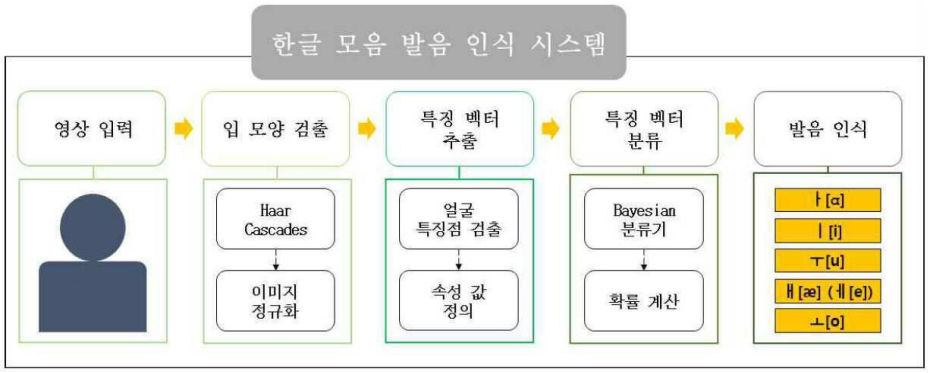
[0261] 실험을 통해서 입 모양 특징 벡터의 확률 분포가 다섯 가지 한글 모음 발음을 구분할 수 있는 모수 분포로 갱신되어지며, 초기 학습 데이터의 양이 적더라도 모음 발음을 인식할 수 있음을 보였다. 결과적으로 학습 데이터를 많이 필요로 하는 CNN을 사용한 유사 시스템보다 좋은 성능을 보였으며, 입 모양에 따른 발음 인식 시스템에 적합하다고 판단된다.

[0262] 실험 결과로 최대 93%의 인식률을 보이며, 이러한 시각 정보만을 활용한 발음 인식 연구는 여러 플랫폼에 적용이 가능하여, 소음이 심한 환경이나 청각적 불편함이 있는 사람들에게 효과적인 인터페이스를 제공할 수 있으며, 자동 자막 생성과 같은 연구에 도움이 될 수 있다.

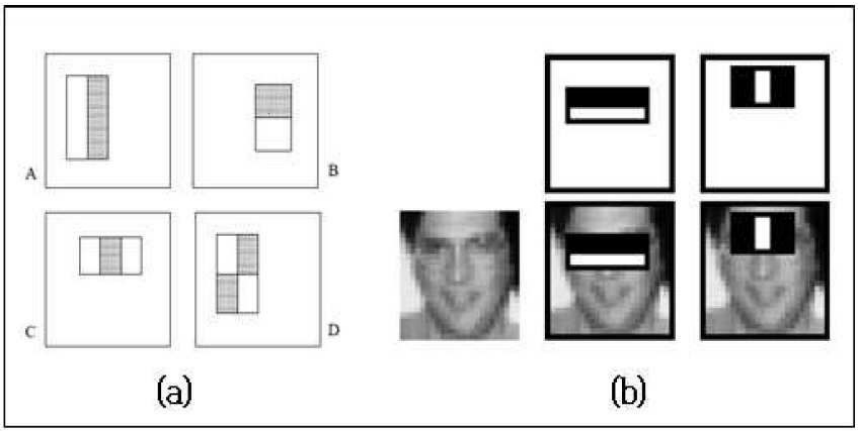
[0264] 이와 같이, 본 발명이 속하는 기술분야의 당업자는 본 발명이 그 기술적 사상이나 필수적 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적인 것이 아닌 것으로서 이해해야만 한다. 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 등가개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면

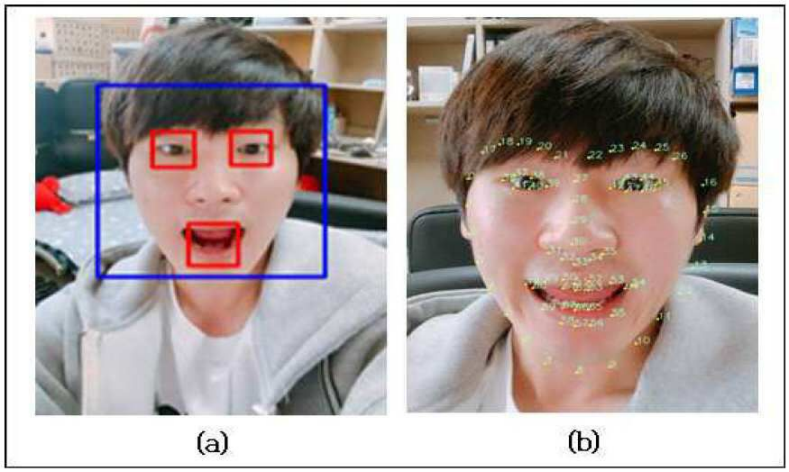
도면1



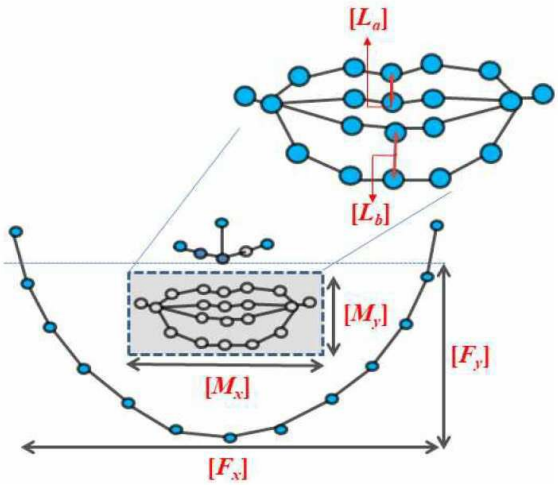
도면2



도면3

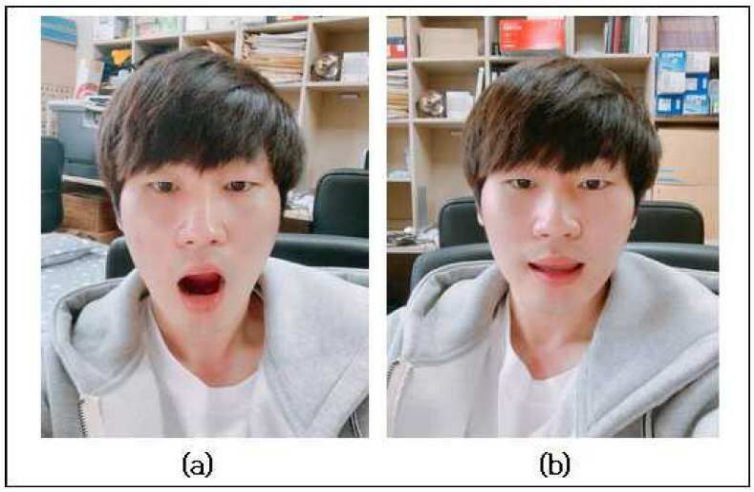


도면4

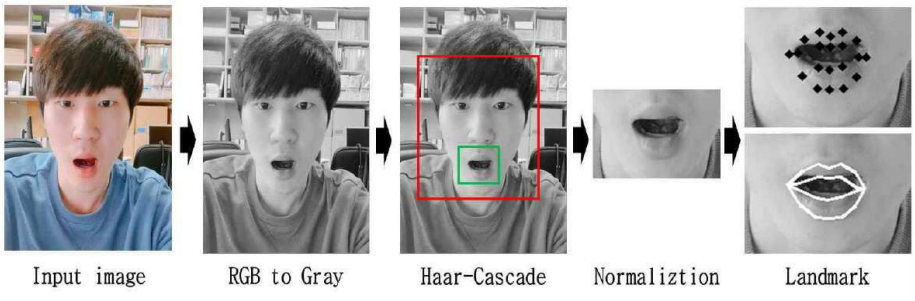


속성 값	M_x	M_y	L_a	L_b	F_x	F_y
의미	입의 수평 길이	입의 수직 길이	윗 입술의 두께	아랫입술의 두께	양쪽 볼의 수평 길이	양쪽 볼의 수직 길이

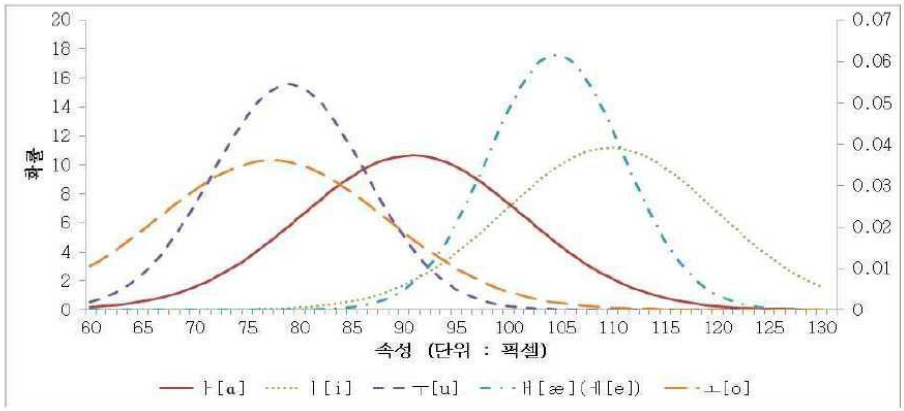
도면5


(a)
(b)

도면6

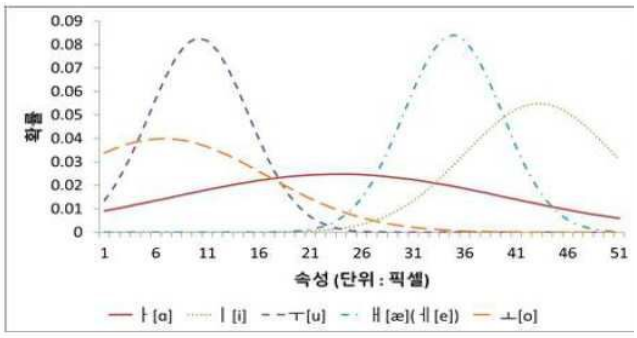

Input image
RGB to Gray
Haar-Cascade
Normaliztion
Landmark

도면7

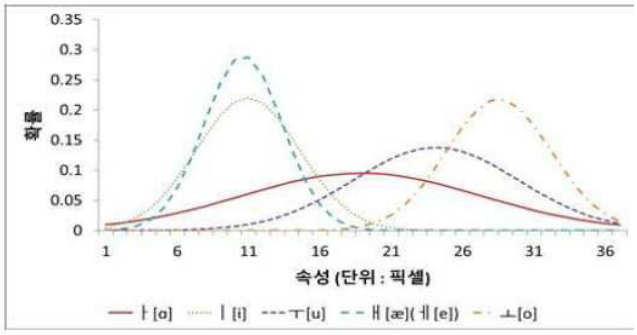

20
18
16
14
12
10
8
6
4
2
0
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
60 65 70 75 80 85 90 95 100 105 110 115 120 125 130
속성 (단위 : 픽셀)
ㅏ[a] ㅣ[i] ㅜ[u] ㅐ[æ](ㅔ[e]) ㅗ[o]

도면8

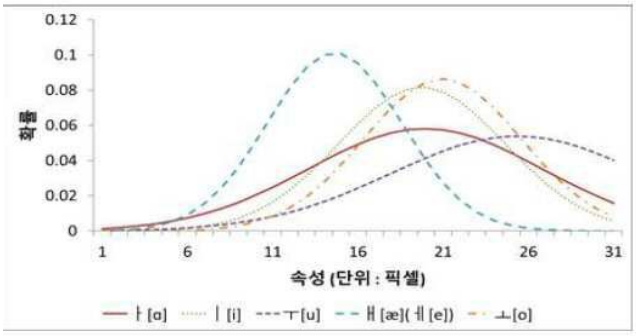
도면9


0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
확률
1 6 11 16 21 26 31 36 41 46 51
속성 (단위 : 픽셀)
ㅏ [a] ㅣ [i] ㅜ[u] ㅐ [æ](ㅔ [e]) ㅗ[o]

(a)

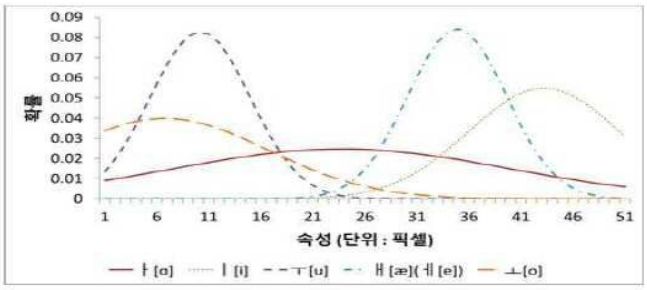

0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
확률
1 6 11 16 21 26 31 36
속성 (단위 : 픽셀)
ㅏ [a] ㅣ [i] ㅜ[u] ㅐ [æ](ㅔ [e]) ㅗ[o]

(b)

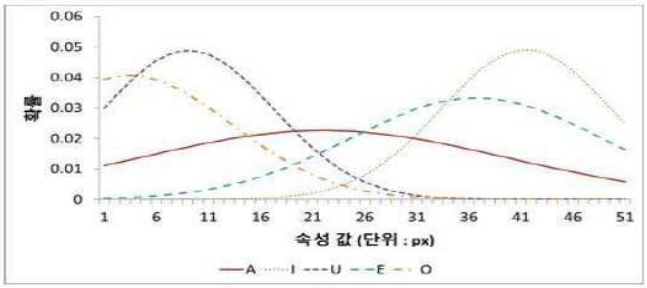

0.12
0.1
0.08
0.06
0.04
0.02
0
확률
1 6 11 16 21 26 31
속성 (단위 : 픽셀)
ㅏ [a] ㅣ [i] ㅜ[u] ㅐ [æ](ㅔ [e]) ㅗ[o]

(c)

도면10

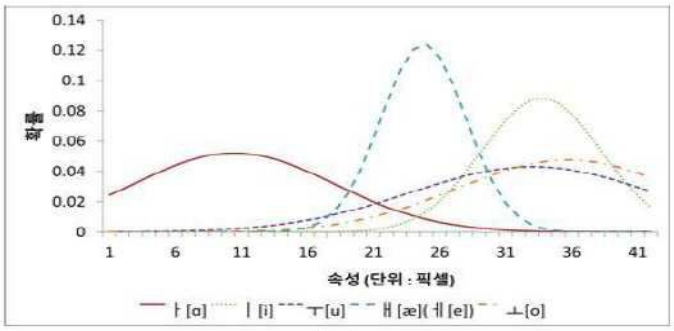


(a)

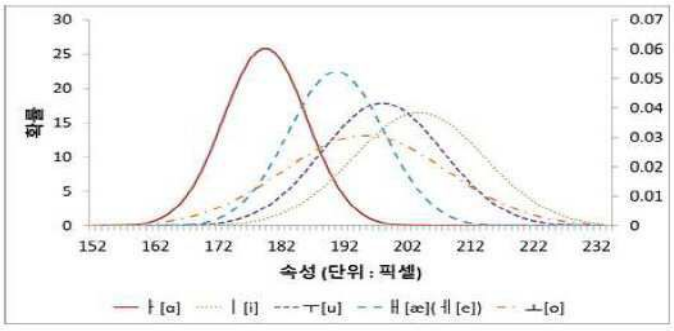


(b)

도면11

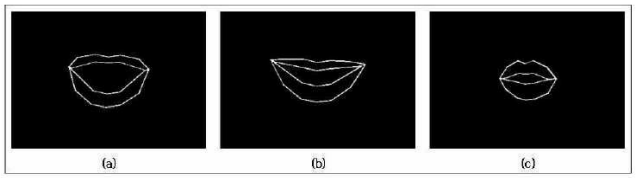


(a)

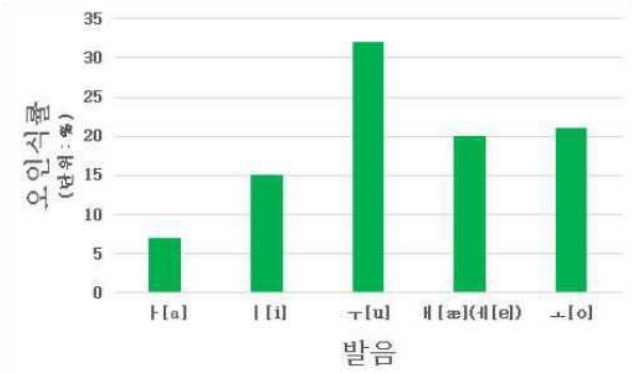


(b)

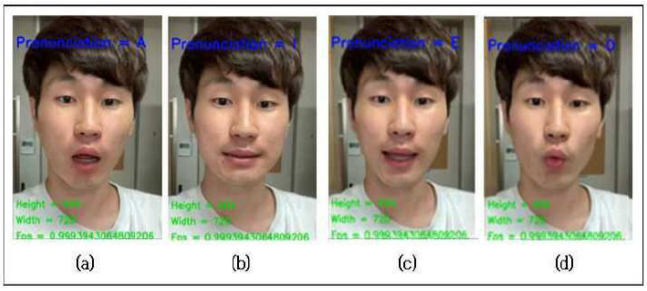
도면12



도면13



도면14



도면15