베이지안 분류 기반 입 모양 인식을 통한 발음 구별법에 대한 연구

김성우*, 차경애, 박세현
대구대학교 정보통신공학과
*e-mail : sungwoo6@daegu.ac.kr

A Study on the Distinguishing Method of Pronunciation through the Recognition of Mouth Shape Using Bayesian Classification

Seong-Woo Kim, Kyung-Ae Cha, Se Hyun Park
*Dept of Telecommunication Engineering, Daegu University

요 약

컴퓨터가 발전함에 따라 실시간 영상에서 입 모양을 검출하여 발음을 구별하는 연구가 진행되고 있다. 이는 음성정보 없이 영상만으로 발음을 구별하여 다양한 어플리케이션에 적용할 수 있는 기술이다. 그러나 현재까지 대부분의 발음 인식 연구는 음성과 영상정보를 동시에 사용하는 방법이 일반적이며, 실시간 영상에서 사람의 입 모양을 검출하는 연구는 부족한 현황이다. 본 논문에서는 베이지안 학습 모델을 기반으로 실시간 영상에서 다섯 가지 발음을 구별하는 시스템을 구현하여 영상만 사용한 발음 구별 연구를 진행하였다.

1. 서론

영상 내에서 사람 얼굴에서 발음 인식을 위해서는 입 모양의 검출이 우선되어야 한다. 이를 위한 방법은 사람의 얼굴에서 눈을 찾고 눈의 위치를 사용하여 입의 위치를 찾는 연구[1], 입술의 밝기 차이를 이용하여 입을 찾는 연구[2] 등 다양한 방법으로 입의 위치를 찾는 연구가 있다. 최근의 입 모양을 통한 문자나 단어 인식기술은 단순히 영상과 음성 정보를 함께 사용하여 두 정보의 교차점으로 영상에서 문자를 찾는 연구이다[3]. 그러나, 발음 인식을 구현하는 많은 부분이 음성 정보에 의존하고 있다.

본 논문에서는 영상 정보만을 사용하여 사람 얼굴에서 입의 위치를 찾고, 입술과 입술 사이의 거리 차이를 이용하여 베이지언 분류기를 통해 입의 모양으로 문자를 식별하게 된다. 입 모양으로 문자를 인식하기 위해서는 첫 번째로 사람 얼굴을 검출하여 입의 위치를 찾고, 두 번째로 찾은 얼굴과 입에 특징 점을 부여하여 입술과 입술, 턱과 입술, 왼쪽 볼과 오른쪽 볼 사이의 위치 차이를 계산한다. 세 번째로 베이지안 분류기에 계산된 위치 차이 값을 애트리뷰트로 사용하여 분류기를 학습시키고 최종적으로 문자를 찾아내어 화면에 나타내게 된다.

2. 관련 연구

2.1 Haar-Cascades classifiers

본 논문에서 사용하고 있는 첫 번째 알고리즘은 얼굴 검출 알고리즘인 Haar-Cascades classifiers 알고리즘이다. 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[4]. 영상에서 해당 영역과 다른 영역간의 밝기 차이를 이용하여 해당 객체를 검출하는 방법으로, 다양한 Feature형태가 존재한다. 그림 1과 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.

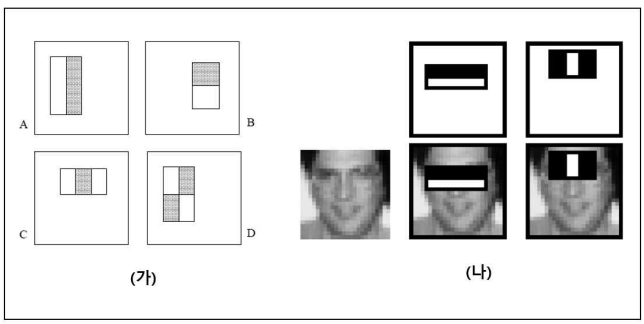

그림 1. 4개의 Haar-like Feature(가)
Haar-like Feature로 영상을 탐색하는 모습(나)

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 영상 내에서 객체를 검출하는 속도를 향상시킬 수 있다는 장점이 있다.

2.2 나이브 베이지안 분류

나이브 베이지안 분류기는 베이즈 이론에 바탕을 두고 있다[5]. 베이즈 이론은 클래스와 애트리뷰트로 나누어져 있고, 하나의 애트리뷰트 값을 기준으로 다른 애트리뷰트들이 독립적이라고 전제로 분류한다. 해당 애트리뷰트 값이 각 클래스에 미치는 영향을 측정하여 애트리뷰트 집합이 어떤 클래스에 속하는가를 분류한다.

예를 들어 애트리뷰트 값 A,B가 있을 때, 클래스1에 속할 확률이 P1(A,B)이고 클래스2에 속할 확률이 P2(A,B)인 경우 P1(A,B)와 P2(A,B) 값을 비교하여 더 높은 값에 애트리뷰트가 해당 클래스에 속한다고 분류한다.

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

그림 2. 베이즈 정리

베이즈 이론은 그림 2와 같은 수식으로 나타낼 수 있는데 여기서 P(X|Y)는 Y사건이 일어난다고 가정했을 때 X가 일어날 확률로 이를 조건부 확률이라고 한다. P(X)와 P(Y)는 각 X사건과 Y사건이 일어날 확률로 사전 확률이라고 하며 서로에 대해 어떤 정보도 가지고 있지 않다는 가정을 전제로 한다. P(Y|X)는 X사건이 일어난다고 가정했을 때의 Y의 조건부 확률이 되며 이를 우도(Likelihood)라고 한다. 나이브 베이즈 분류기에서는 사전확률인 P(X), P(Y)와 우도(Likelihood) 값을 통해 최대가 되는 조건부 확률 값을 찾는다. P(X|Y) 최대가 되는 조건부 확률 값은 애티리뷰트의 튜플이 해당 클래스에 속할 가능성이 높다는 것이 된다.

본 논문에서는 'A', 'I', 'U', 'E', 'O'발음을 각각 하나의 클래스로 정의하고, 얼굴에서 추출한 특징 벡터를 애트리뷰트로 정의하여 나이브 베이지안을 통해 입 모양을 분류할 수 있다.

3. 베이지안 분류를 이용한 발음 구별

본 논문은 Haar Cascades 알고리즘을 사용하여 사람의 정면 얼굴을 검출한 다음, dlib[6]에 정의되어있는 얼굴 특징점들을 이용하여 찾은 얼굴의 각 특징점들의 차이 값을 베이지안 분류기에 적용하였다.

베이지안 분류기에 사용된 애트리뷰트는 총 5가지로 (1) 입의 왼쪽 끝점과 오른쪽 끝점의 거리[Ma], (2)위 입술의 두께[Mb], (3)아래 입술의 두께[Mc], (4)위 입술과 아래 입술까지의 거리[Md], (5)위 입술에서 턱까지의 거리[Me], (6)왼쪽 볼과 오른쪽 볼 사이의 거리[Mf]로 설정하였다.

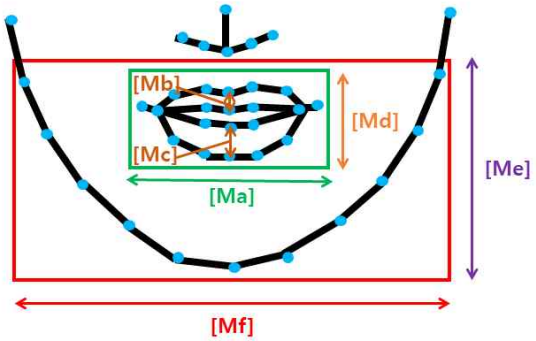

그림 3. 얼굴 특징점과 애트리뷰트

입 모양에 따라 각 (2)과 (3)은 달라지고, 입을 많이 벌릴수록 (4)와 (5)는 커지게 된다[7]. 또한 'A'발음에서 'I'나 'E'발음으로 변할수록 양 볼이 넓어진다는 점을 볼 때, 입과 입술 크기가 비슷한 'A'와 'E'의 큰 차이점으로 (6)는 'A'와 'E'를 구별할 수 있는 애트리뷰트가 된다. 이를 바탕으로 기존의 연구들과 유사하지만 음성을 사용하지 않고 영상 하나만을 이용하여 입 모양을 실시간으로 정확하게 분류하는 것이 본 논문의 목표이다.

4. 개발 및 실험 결과

제안하는 입 모양 검출 시스템은 Python과 OpenCV로 구현하였다. 연구에서 사용된 영상데이터로는 20대 후반의 남,여와 30대 초반의 남, 여 그리고 50대 후반의 남,여 이미지 영상을 사용하여 총 500개의 영상 데이터를 사용하였다. 영상 데이터는 'A', 'I', 'U', 'E', 'O' 각각 100개씩 5개 발음의 정면 얼굴로 구성되어 있다. 실시간 영상에서는 사람이 영상에 비치는 거리에 따라 해당 애트리뷰트 값이 변하기 쉬우므로 이미지 크기 변환이 필수이다. 시스템을 구현하기 위해서는 먼저 입력 영상에서 Haar cascades classifier 알고리즘으로 사람 얼굴 영역을 찾는다. 이후 작업에서는 전체 영상을 사용하지 않고, 잘라낸 이미지로 얼굴에서 각 특징점들을 찾고, 특징점들 사이의 거리를 계산한다. 이렇게 함으로써 영상에서의 거리에 관계없이 사람 얼굴을 일정한 크기의 영상으로 만들어 애트리뷰트 값을 비슷하게 얻을 수 있다. 다음으로는 받아온 애트리뷰트 값들을 하나의 튜플로 저장한 다음, 저장된 튜플에 해당 클래스 값을 부여한다. 이렇게 600장의 이미지 데이터에서 해당 이미지의 애트리뷰트 튜플과 클래스 데이터를 수집하게 되고 이 데이터들을 학습 데이터로 사용하여 나이브 베이지안 분류기에 적용한다.

테스트 데이터로는 확장자가 MPEG-4 동영상을 사용하였으며 각 프레임마다 학습 데이터와 같은 방법으로 애트리뷰트를 추출하게 되고, 여기서 학습 데이터를 사용하여 해당 애트리뷰트 값이 어떤 클래스에 속하는지 나이브 베

이지안 분류기로 분류하여 해당 클래스를 판단하게 된다. 아래의 표는 클래스별로 추출한 5개의 이미지에서의 애트리뷰트와 클래스 값을 나타내고 있다.

표 1. 'A'발음에 대한 애트리뷰트 측정값 (단위 : px)

	Ma	Mb	Mc	Md	Me	Mf
A1	87.9	91.7	21.9	72.8	120.4	191.1
A2	85.8	96.4	22.7	91.3	139.4	180.9
A3	89.3	99.2	24.1	85.6	133	185
A4	89.3	99.2	24.1	85.6	133	185
A5	89.3	99.2	24.1	85.6	133	185

표 2. 'I'발음에 대한 애트리뷰트 측정값 (단위 : px)

	Ma	Mb	Mc	Md	Me	Mf
I1	114.6	113.4	22.7	46.2	115.5	223.3
I2	115.1	110.2	21.3	43.9	105.8	221.9
I3	116	117	22.7	46.1	110.5	223.9
I4	111.7	110	22.6	44.6	119.7	214.8
I5	114.4	110.5	22.6	46.7	119.6	222.6

표 3. 'U'발음에 대한 애트리뷰트 측정값 (단위 : px)

	Ma	Mb	Mc	Md	Me	Mf
U1	84.4	103.4	24.8	39.6	95.6	204.7
U2	84.4	106.1	26.2	41.7	97.6	206.5
U3	83.7	100.4	26.2	40.3	98.4	200.1
U4	82.3	104.1	26.9	42.4	99.7	203.5
U5	79.4	98.3	26.9	43.1	102.6	197.7

표 4. 'E'발음에 대한 애트리뷰트 측정값 (단위 : px)

	Ma	Mb	Mc	Md	Me	Mf
E1	79.9	100.7	29	46.7	103.1	191.3
E2	115.3	111.7	24.1	55.9	121.3	206.5
E3	115.3	111.7	24.1	55.9	121.3	206.5
E4	116.7	116.2	25.5	55.9	130.2	207.9
E5	120.9	108.3	23.3	58	124.5	203

표 5. 'O'발음에 대한 애트리뷰트 측정값 (단위 : px)

	Ma	Mb	Mc	Md	Me	Mf
O1	70.9	108.1	26.9	55.9	126.6	212.8
O2	70.9	108.1	26.9	55.9	126.6	212.8
O3	88.5	130.5	38.9	78.6	138.3	243
O4	84.1	112.9	29.7	60.8	149.9	223.5
O5	71.4	95.2	29	62.9	112.4	192.5

위의 표 1을 보면 'A' 발음은 다른 발음에 비해 입의 크기 값인 'Md'의 크기가 크다는 것을 알 수 있다. 'I' 발음은 입술의 두께인 'Mb', 'Mc'의 값이 작고 볼과 볼까지의 길이 값인 'Mf'가 크다는 특징이 있는데, 사람이 'I'발음을 할 때는 양 볼이 넓어지기 때문이다. 비슷한 입 모양인 'U'와 'O'의 큰 차이점은 입과 턱 사이의 길이인'Me'와 양 볼 사이의 길이인 'Mf'이다. 'U'발음보다 'O'발음을 할 때 사람은 입을 벌리게 되므로 'Me'의 크기가 훨씬 커지게 된다. 그리고 'E'발음은 'A'와 'I'발음의 중간 값을 띄고 있어 오 인식을 불러오는 계기가 된다. 'E'발음은 가장 검출이 어렵기 때문에 많은 학습 데이터를 필요로 한다.

그림 4는 실시간 영상의 각 프레임에서 발음을 검출하여 영상에 나타낸 사진이다.

그림 4. 실시간 영상에서의 발음이 검출 결과

그림 4에서는 1초에 30frame의 30초짜리 'A′, 'I′, 'U′, 'E′, 'O′ 발음 영상을 사용하였다. 1frame마다 입 모양을 검출하여 애트리뷰트 값을 받아왔고 총 900개의 애트리뷰트 튜플을 얻을 수 있었다. 결과적으로 발음에 따라 정확한 입모양 검출이 가능하였다.

5. 결론

본 논문에서는 기존의 영상 이미지와 음성을 사용하여 영상에서의 발음을 검출하는 기존 연구와 달리 영상 이미지를 이용하여 실시간으로 사람 얼굴과 입을 검출하고, 입 모양에 따른 발음 클래스를 분류하는 시스템을 구현하였다. 'A′, 'I′, 'U′, 'E′, 'O′ 발음은 각각 다른 입 모양을 가지고 있으므로 영상에서의 검출이 용이하다는 장점이 있다. 사람 얼굴을 검출하기 위하여 Haar cascades classifier 알고리즘을 사용하여 사람 얼굴을 검출하고, 각 얼굴의 특징점을 이용하여 각 특징점 사이의 거리를 애트리뷰트로 지정하였다. 그리고 해당 애트리뷰트 값을 이용하여 각 발음 클래스에 대해 나이브 베이지안으로 분류하여 실시간 영상에서 발음을 검출하였다. 단점으로는 영상의 밝기와 얼굴 각도에 따라 정확도가 달라지고, 얼굴이 가려지거나 정면을 보지 않을 경우 검출이 힘들다는 점이 있다. 본 논문에서는 다양한 사람들의 영상을 학습한 데이터로 한 사람의 입 모양을 검출하는 시스템을 구현하였다. 이후에는 영상에서 검출된 사람 얼굴을 데이터로 사용하여 사

람에게 ID를 부여하고 각 사람에 따른 정확한 입모양 발음을 학습하여 각 사람에 대해 더욱 정확한 입모양을 검출하는 기술에 대한 연구를 진행할 예정이다.

참고문헌

[1] 송민규, Thanh Trung Pham, 김진영, 황성택, "모바일 환경에서의 시각 음성인식을 위한 눈 정위 기반 입술 탐지에 대한 연구," 한국지능시스템학회 논문지, 제19권, 4호, pp. 478-484, 2009.

[2] 김기백, 유제웅, 조남익, "입술 영역의 움직임과 밝기 변화를 이용한 음성구간 검출 알고리즘 개발," 방송공학회 논문지, 제17권, 3호, pp. 519-528, 2012.

[3] 임대영, 김선광, 정길도, "영어발음 향상을 위한 실시간 인공지능 입모양 인식 프로그램 개발," 제어로봇시스템학회 논문지, 제24권, 4호, pp. 327-333, 2018.

[4] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features," *CVPR*, 1, No. 1, pp. 511-518, 2001.

[5]Andrew R. Webb., Statistical pattern recognition, Wiley, (2003 printing).

[6] Dlib C++ Library, http://dlib.net/

[7] 이근민, 한경임, 박혜정, "SVM 기법에 기초한 청각장애인의 영어모음 발음을 위한 음성 인식 및 입술형태 특징 추출," 재활복지공학회논문지, 제11권, 3호, pp. 247-252, 2017.