정규논문 (Regular Paper)
방송공학회논문지 제17권 제3호, 2012년 5월 (JBE Vol. 17, No. 3, May 2012)
http://dx.doi.org/10.5909/JBE.2012.17.3.519

입술 영역의 움직임과 밝기 변화를 이용한 음성구간 검출 알고리즘 개발

김 기 백[a], 유 제 웅[b], 조 남 익[b]‡

Voice Activity Detection using Motion and Variation of Intensity in The Mouth Region

Gibak Kim[a], Jewoong Ryu[b], and Nam Ik Cho[b]‡

요 약

음성구간을 검출하는 일반적인 방법은 음향신호로부터 특징값을 추출하여 판별식을 거치는 것이다. 그러나 잡음이 많은 환경에서 그 성능은 당연히 저하되며, 이 경우 영상신호를 이용하거나 영상과 음성을 동시에 사용함으로써 성능향상을 도모할 수 있다. 영상신호를 이용하여 음성구간을 검출하는 기존 방법들에서는 액티브 어피어런스 모델, 옵티컬 플로우, 밝기 변화 등 주로 하나의 특징값을 이용하고 있다. 그러나 음성구간의 참값은 음향신호에 의해 결정되므로 한 가지의 영상정보만으로는 음성구간을 검출하는데 한계를 보이고 있다. 본 논문에서는 입술 영역의 옵티컬 플로우와 밝기 변화 두 가지 영상정보로부터 특징값을 추출하고, 추출된 특징값들을 결합하여 음성구간을 검출하는 알고리즘을 제안하고자 한다. 또한, 음성구간 검출 알고리즘이 다른 시스템의 전처리로 활용되는 경우에 적은 계산량만으로 수행되는 것이 바람직하므로, 통계적 모델링에 의한 방법보다는 추출된 특징값으로부터 간단한 대수적 연산만으로 스코어를 산정하여 문턱값과 비교하는 방법을 제안하고자 한다. 입술 영역 검출을 위해서는 얼굴에서 가장 두드러진 특징점을 갖는 눈을 먼저 검출한 후, 얼굴의 구조와 밝기값을 이용하는 알고리즘을 제안하였다. 실험 결과 본 논문에서 제안하는 두 가지 특징값을 결합한 음성구간 검출 알고리즘이 하나의 특징값만을 이용했을 때보다 우수한 성능을 보임을 확인할 수 있다.

Abstract

Voice activity detection (VAD) is generally conducted by extracting features from the acoustic signal and a decision rule. The performance of such VAD algorithms driven by the input acoustic signal highly depends on the acoustic noise. When video signals are available as well, the performance of VAD can be enhanced by using the visual information which is not affected by the acoustic noise. Previous visual VAD algorithms usually use single visual feature to detect the lip activity, such as active appearance models, optical flow or intensity variation. Based on the analysis of the weakness of each feature, we propose to combine intensity change measure and the optical flow in the mouth region, which can compensate for each other's weakness. In order to minimize the computational complexity, we develop simple measures that avoid statistical estimation or modeling. Specifically, the optical flow is the averaged motion vector of some grid regions and the intensity variation is detected by simple thresholding. To extract the mouth region, we propose a simple algorithm which first detects two eyes and uses the profile of intensity to detect the center of mouth. Experiments show that the proposed combination of two simple measures show higher detection rates for the given false positive rate than the methods that use a single feature.

Keyword : Visual voice activity detection, Optical flow, Intensity variation

Ⅰ. 서 론

음성구간 검출은 입력 음향신호에서 음성구간과 비음성구간을 구별하여 검출하는 것으로서, 음성인식, 음성압축, 잡음제거 등 여러 응용분야에서 필수적인 전처리 과정이다. 음성압축에서는 음성구간 검출을 통해 비음성구간에는 낮는 비트율을 할당하여 대역폭을 줄일 수 있다[1,2]. 잡음이 섞인 신호에서 잡음을 제거하기 위해서는 먼저 잡음의 특성을 구해야하는 경우가 많은데, 음성구간 검출기를 이용하여 비음성구간에서 잡음의 특성을 구하는 방식이 널리 사용되고 있다[3]. 음성인식 시스템에서는 최근 들어 비음성 음향모델을 이용하여 음성구간 검출기의 도움이 필요 없는 경우도 있으나, 계산량 절감 및 성능 향상을 위하여 인식대상 발화의 시작점과 끝점을 검출하여 검출된 발음에 대해서만 음성인식 과정을 수행하는 방식이 여전히 많이 사용되고 있어 음성구간 검출기의 성능이 음성인식 성능에 큰 영향을 미치고 있다[4].

일반적으로 음성구간 검출은 입력 음향신호로부터 추출된 특징값을 판별식에 적용하여 이루어지므로 입력 음향환경에 많은 영향을 받게 된다. 잡음이 심한 환경에서는 음성구간 검출기의 신뢰도가 많이 떨어지는데, 이를 해결하고자 잡음에 강인한 특징값을 사용하거나 하나 이상의 마이크를 사용하여 공간정보를 이용하는 등의 노력이 이루어지고 있다[5-7]. 그러나 잡음이 음성과 비슷한 스펙트럼을 갖거나 그 특성이 비정상적 (non-stationary)인 경우, 또는 신호대 잡음비가 매우 낮은 경우에는 여전히 한계를 보이고 있다. 따라서 잡음에 영향을 받지 않는 영상신호만을 이용하거나 음향신호와 영상신호를 함께 사용함으로써 음성구간 검출 성능을 향상시키고자 하는 연구도 수행되어 왔다.

영상신호를 이용하는 연구는 주로 입술의 움직임을 이용하는 것이며[8], 음성과 영상을 함께 이용하는 멀티모달 시스템이 점차적으로 확산됨에 따라 음향잡음이 심한 환경에서 영상신호를 이용하여 음성구간 검출 성능을 향상시키려는 시도도 많이 이루어지고 있다[9-14]. 영상신호를 이용한 음성구간 검출 알고리즘은 특징값을 추출하는 방법과 추출한 특징값을 이용하여 음성 비음성을 판별하는 방법에 따라 여러 방식의 알고리즘이 제안되어 왔다. Liu와 Wang은 입술 주변 영역을 주성분분해법 (PCA: Principal Component Analysis)으로 구해진 주성분공간으로 투영한 후, 비음성구간은 하나의 가우시안 분포로 모델링하고 음성구간은 가우시안 믹스처 분포로 모델링 (GMM: Gaussian Mixture Model)하여 음성구간을 검출하였다[9]. Aubrey 등은 화자의 입술로부터 추출된 어피어런스 (appearance) 파라메터[10]를 은닉 마코프 모델 (HMM: Hidden Markov Models)로 모델링하는 방법을 사용하여 음성구간을 검출하였다[11]. Siatras 등은 발음할 때 입이 벌어지면서 어둡게 보이는 구강이 많이 나타난다는 사실에 착안하여 입술영역에서 문턱값보다 어두운 화소들의 개수를 이용하여 음성구간 검출을 시도하였다[12]. 발화가 이루어지는 경우에는 입술영역의 어두운 화소의 개수가 증가 또는 많은 변화가 발생하게 된다. 이러한 영상정보가 음성과 비음성구간에서 각각 가우시안 분포를 따른다고 가정하여 모델링하는 방법을 사용하였다. Navarathna 등은 영상정보를 이용한 음성인식에서 사용되었던 일련의 변환법을 사용하여 음성구간 검출에 적용하였는데[13], 여기서는 정적 특징값 뿐만 아니라 동적 특징값을 사용하였고 가우시안 믹스처 모델링을 통해 음성/비음성 구간을 판별하였다. 또한, Aubrey 등은 입술의 움직임을 이용하여 음성구간을 검출하려고 시도하였는데, 각 영상 프레임에서 구한 옵티컬 플로우의 변화를 HMM을 이용하여 모델링하였다[14].

이와 같이 영상기반의 음성구간 검출은 다양한 특징값을 이용하여 왔으나 음성구간의 참값은 음향신호에 의해 결정되므로 한 가지의 영상정보만으로는 음성구간을 검출하는데 한계를 보이고 있다. 예를 들어 문장 끝에 위치하는 모음과 같이 입술 움직임이 없이 음성이 지속되는 구간에서 음성구간 검출을 하는 경우, 만일 입술 움직임으로부터 특징

a) 숭실대학교 전기공학부 (School of Electrical Engineering, Soongsil University)
b) 서울대학교 전기컴퓨터공학부 (Department of Electrical Engineering and Computer Science, Seoul National University)
‡ 교신저자 : 조남익 (Nam Ik Cho)
E-mail: nicho@snu.ac.kr
Tel: +82-2-880-1810, Fax: +82-2-880-8183
※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행되었습니다 (2012-0003455).
・접수일(2012년3월28일),수정일(2012년4월27일),게재확정일(2012년4월27일)

값을 추출하여 음성구간을 검출한다면 문장 끝에 위치하는 모음의 끝부분은 비음성으로 간주될 위험이 높다. 이러한 경우는 어두운 구강영역이 증가되어 있다는 사실을 이용하면 어느 정도 검출오류를 줄일 수 있을 것이다. 이와 같이 하나 이상의 영상정보 특징값을 이용한다면 상호보완 효과가 기대되므로 음성구간 검출 성능을 높일 수 있을 것이다. 따라서 본 논문에서는 입술 영역의 옵티컬 플로우와 밝기 변화 등 두 가지 영상정보로부터 특징값을 추출하는 방법을 제안하고, 추출한 특징값들을 결합하여 음성구간을 검출하는 알고리즘을 제안한다. 또한, 음성구간 검출 알고리즘이 다른 시스템의 전처리로 활용되는 경우에는 적은 계산량만으로 수행되는 것이 바람직하므로, 통계적 모델링에 의한 방법보다는 추출된 특징값으로부터 간단한 대수적 연산만으로 스코어를 산정하여 문턱값과 비교하는 방법을 제안한다.

제안하는 영상정보 기반의 음성구간 검출기는 입술 주변 영역의 움직임과 픽셀 밝기 변화를 이용하기 때문에 입술 영역을 검출하는 과정이 선행되어야 한다. 본 논문에서는 입술을 검출하는 것이 아니라 입술을 포함한 영역을 대략적으로 검출하면 되므로 기존의 복잡한 방법 대신 적은 계산량으로 간단하게 구현할 수 있는 방법을 제안한다. 제안하는 방법은 얼굴의 가장 두드러진 특징점인 두 눈을 찾은 후, 얼굴의 일반적인 형태와 밝기 정보를 이용하여 입술 영역을 찾아내는 방법이다. 화자의 발성이 담긴 동영상에 대한 실험 결과에서 본 논문에서 제안하는 바와 같이 두 가지 특징값을 결합한 음성구간 검출 알고리즘이 하나의 특징값만을 이용했을 때보다 우수한 성능을 보임을 확인할 수 있다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 음성구간 검출을 위해 사용하는 영상정보 기반의 특징값에 대해 설명하고, 3장에서는 두 가지 특징값을 결합하여 음성구간을 검출하는 과정을 보여준다. 4장에서는 얼굴과 입술 영역 검출 과정에 대해 기술한다. 영상 데이터를 이용한 실험결과는 5장에서 제시한다.

II. 음성구간 검출을 위한 특징값 추출

본 장에서는 음성구간 검출을 위한 두 가지 영상정보 기반 특징값 추출에 대해 설명한다. 옵티컬 플로우를 이용한 특징값은 입술과 그 주변영역의 움직임을 이용하는 것이며 밝기 정보를 이용한 특징값은 발성시 변화하는 구강영역의 밝기 변화를 이용하는 것이다.

1. 옵티컬 플로우를 이용한 특징값 추출

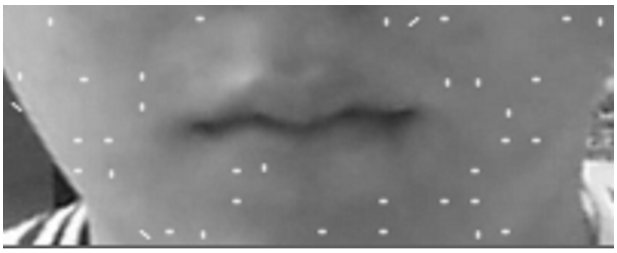
사람이 발성할 때의 영상에서 가장 특징적인 것은 입술과 그 주변 영역의 움직임이다. 즉, 연속된 영상들에서 나타나는 입술 영역의 움직임이 화자가 말하고 있음을 알려주는 신호가 될 수 있다. 이러한 입술 영역의 움직임을 포착하기 위해 옵티컬 플로우 (optical flow)를 사용할 수 있다[14,15]. 옵티컬 플로우는 영상의 움직임을 나타내는 일종의 패턴으로서 연속된 프레임에서 나타나는 픽셀의 이동을 속도벡터로 나타낸다. 본 논문에서는 옵티컬 플로우 추정에 널리 사용되는 Lucas-Kanade 방법을 사용하였다[16,17]. 옵티컬 플로우로 얻어진 속도 벡터는 픽셀 단위로 얻어진 것이기 때문에 영상 잡음에 취약하다. 따라서 본 논문에서는 아래의 식과 같이 $M\times M$ 크기의 격자로 나누어서 각 격자에 해당하는 옵티컬 플로우의 평균값으로 입술 영역의 움직임을 격자 속도벡터($\mathbf{v}_{\mathrm{grid}}(i,j)$)로 나타내도록 하였다.

$$\mathbf{v}_{\mathrm{grid}}(i,j)=\frac{\sum_{(x,y)\in C_{ij}}\mathbf{v}_{\mathrm{pixel}}(x,y)}{M^2}. \tag{1}$$

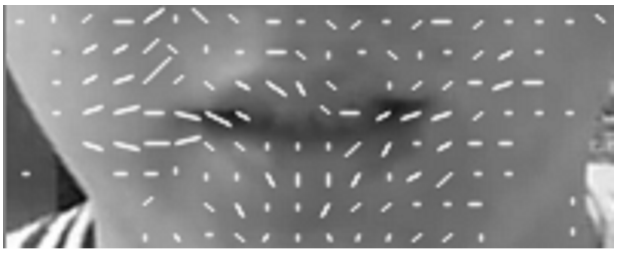
여기서 $\mathbf{v}_{\mathrm{grid}}(i,j)$ 와 $\mathbf{v}_{\mathrm{pixel}}(x,y)$는 각각 격자 (i,j)에서의 격자 단위 옵티컬 플로우와 픽셀 (x,y)에서의 픽셀 단위 옵티컬 플로우를 나타내며 C_{ij}는 $M\times M$ 크기의 격자 (i,j) 내부에 있는 픽셀들을 원소로 하는 집합을 의미한다. 이 때, 광역 움직임 (global motion) 즉, 얼굴전체의 움직임에 의한 영향을 보상하기 위하여 격자 단위 옵티컬 플로우의 평균값을 차감하였다.

그림 1에 격자 속도벡터의 예시영상을 나타내었다. 그림 1을 보면 발성구간에서의 영상으로부터 추출한 속도벡터의 크기가 묵음구간에 비해 크게 나타나고, 방향은 일관되지 않게 나타나는 것을 알 수 있다. 본 논문에서는 음성구간

검출을 위해 속도벡터의 방향은 고려하지 않고, 각 영상 프레임에서 다음 식과 같이 아래쪽 얼굴 영역의 속도 벡터 크기 중 가장 큰 값을 특징값으로 이용한다.



(a) 묵음구간에서의 옵티컬 플로우



(b) 발음하고 있을 때의 옵티컬 플로우

그림 1. 입술과 주변 영역에서의 옵티컬 플로우 (흰색 화살표는 속도 벡터를 나타낸다.)
Fig. 1. The optical flow of the mouth region

$$F_o[n] = \max_{(i,j)\in C_{ij}^{Mouth}} |\mathbf{v}_{\mathrm{grid}}(i,j)| \quad . \tag{2}$$

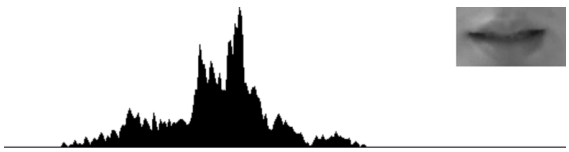
여기서 n은 프레임 인덱스이며 C_{ij}^{Mouth}는 입술 영역으로 검출된 영역의 격자 (i,j) 내부에 있는 픽셀들을 원소로 하는 집합을 의미한다.

2. 밝기 정보를 이용한 특징값 추출

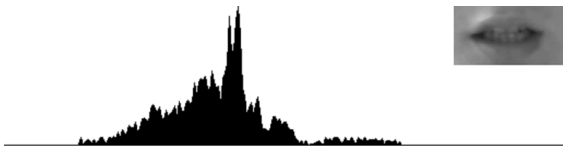
본 논문에서 음성구간 검출에 사용하고자 하는 또 다른 영상정보는 입술 영역의 밝기 정보이다. 일반적으로 말을 하지 않을 때는 일반적으로 입이 닫혀있고, 발성이 이루어질 때는 자연스럽게 입을 벌리게 되어 닫혀있던 구강 (oral cavity)이 드러나게 된다. 그런데 구강은 일반적인 상황에서는 외부의 빛이 닿지 않기 때문에 그늘지게 되어 주변보다 어두운 밝기값을 가진다. Siatras 등의 논문에서는 이러한 사실을 이용하여 문턱값을 설정하고 그보다 어두운 밝기를 갖는 픽셀의 수를 특징값으로 사용하여 이러한 특징값의 증가 및 변화를 통계적 모델링을 통해 음성구간을 검출하였다[12].

그러나 발화 시 어두운 밝기를 가지는 픽셀의 개수만 증가하는 것이 아니라, 치아가 드러나는 경우('이', '의' 등의 모음)에는 주변보다 밝은 밝기를 가지는 픽셀의 개수가 증가하기도 한다. 이러한 관측을 바탕으로 본 논문에서는 픽셀의 밝기가 주변보다 낮거나 높은 픽셀의 개수를 특징값으로 사용한다. 해당되는 픽셀의 개수를 산정하기 위해서는 밝은 픽셀을 검출하기 위한 문턱값과 어두운 픽셀을 세기 위한 문턱값을 각각 설정하여야 한다. 본 논문에서는 두 개의 문턱값 (T_L, T_H)을 다음과 같이 설정한다. 앞 절에서 구한 입술 영역의 속도 정보를 이용하여 움직임이 거의 없을 때 (비음성 구간일 때), 입술영역에 대한 밝기값들로부터 히스토그램을 구하고 상위 1%에 해당하는 밝기값을 T_H로 설정하고 하위 5%에 해당하는 밝기값을 T_L로 설정한다. 여기서 문턱값 설정을 위한 1%와 5%는 실험적으로 구한 값이다. 입술과

(a) 묵음구간



(b) 치아가 감춰진 상태로 발성하는 경우



(c) 치아가 드러나면서 발성하는 경우

그림 2. 세 가지 경우에 대한 밝기의 히스토그램
Fig. 2. Intensity histograms for three different cases

그 주변 영역에서 밝기가 T_L보다 어둡거나 T_H보다 밝은 픽셀의 수를 세어 두 가지 픽셀 수의 합의 변화가 큰 구간을 발성이 이루어지는 구간으로 간주한다. 해당 픽셀의 수 변화를 파악하기 위해, 현재의 프레임을 포함한 이전 N개의 프레임동안 밝거나 어두운 픽셀 수의 표준편차를 구하여 음성구간 검출을 위한 특징값($F_i[n]$, 프레임 인덱스 n에 대하여)으로 사용한다. 본 실험에서 N은 8로 설정하였다. 그림 2에 세 가지 경우에 대한 밝기의 히스토그램의 예를 나타내었다. 묵음 구간 (a)의 히스토그램과 (b), (c)의 히스토그램을 비교하면 문턱값보다 어두운 밝기를 가지는 픽셀의 수는 두 경우 모두 증가하고, 치아가 드러났을 경우에는 밝은 값을 가지는 픽셀의 수도 역시 증가하였음을 알 수 있다.

III. 음성구간 검출 과정

이번 장에서는 앞에서 구한 두 가지 특징값들을 이용하여 음성 구간을 검출하는 과정을 설명한다. 특징값을 이용하여 음성구간을 검출하기 위해 기존에는 GMM 및 HMM 등 통계적인 모델링을 이용한 방법들이 주로 사용되어 왔다[9,11-14]. 본 논문에서는 대규모 데이터를 이용한 훈련과정을 필요로 하지 않고 비교적 간단한 대수적 계산만으로 음성구간을 검출하는 방법을 제안하고자 한다. 먼저 각각의 2장에서 설명한 특징값 F_o와 F_i에 대해 다음식과 같이 음성구간 검출을 위한 스코어를 계산한다.

$$S_{o|i}[n]=\begin{cases}1.0 & ,F_{o|i}[n]>T_p\\ \dfrac{1}{T_p-T_a}(F_{o|i}[n]-T_a) & ,T_a<F_{o|i}[n]\le T_p\\ 0 & ,F_{o|i}[n]\le T_a\end{cases} \quad (3)$$

여기서 $F_{o|i}[n]$과 $S_{o|i}[n]$은 옵티컬 플로우와 밝기 정보에 기반한 방법에 대한 n번째 프레임에서의 특징값과 스코어를 나타낸다. 각 특징값에 대한 스코어는 위 식에서 나타난 바와 같이, 문턱값 T_p보다 큰 경우는 1로 두고, 문턱값 T_a보다 작은 경우는 0으로 두며, 그 사이의 특징값에 해당하는 경우는 선형적으로 변화하도록 설정한다. T_p와 T_a는 각각 발성이 이루어지는 경우와 그렇지 않은 경우에 대한 문턱값으로서 실험적으로 설정되며, 본 논문에서는 옵티컬 플로우 기반의 특징값에 대해서는 각각 2.5와 1.5로 설정되었고, 밝기 정보 기반의 특징값에 대해서는 각각 25와 17로 설정되었다. 이러한 문턱값들은 검출된 얼굴영상의 크기가 비슷하다면 영상에 따라 큰 차이를 갖지 않고 적용될 수 있다. 이와 같이 각 특징값에 대한 스코어를 계산한 후, 아래의 식과 같이 두 스코어의 가중합 (weighted sum)으로 최종스코어를 결정한다.

$$S_c[n]=\alpha\cdot S_o[n]+(1-\alpha)\cdot S_i[n]. \quad (4)$$

본 논문에서 가중치 α는 0.5로 두어 두 특징값의 영향을 동일하게 적용하였다. 하지만 영상의 환경에 따라 가중치에 차등을 둘 수도 있다. 예를 들어, 영상의 밝기가 일반적이지 않아 발성시 드러나는 구강과 치아에 의한 밝기 구분이 어려운 경우는 가중치 α를 0.5보다 큰 값으로 두어 옵티컬 플로우에 의한 스코어에 좀 더 의존하여 성능을 향상시킬 수 있을 것으로 기대한다. 본 논문에서는 차등가중치에 대한 내용은 다루지 않으며 이에 대한 내용은 향후 연구에서 다루고자 한다. 음성/비음성 구별은 프레임 간에 서로 어느 정도의 상관관계가 있고 이를 반영하기 위해 다음 식과 같이 망각인자 (β)가 있는 일차 재귀시스템을 최종 스코어에 적용한다.

$$S_{final}[n]=\beta\cdot S_c[n]+(1-\beta)\cdot S_c[n-1]. \quad (5)$$

본 논문의 실험에서는 β를 0.7로 두었다. 이렇게 구해진 최종스코어를 아래와 같이 어떤 문턱값과 비교하여 현재 프레임이 음성구간인지 비음성구간인지를 구별하게 된다.

$$\text{음성/비음성 검출결과}=\begin{cases}\text{음성구간} & ,\text{if}\, S_{final}[n]>T\\ \text{비음성구간} & ,\text{else}\end{cases} \quad (6)$$

T는 문턱값으로서 T가 크면 정검출율 (true positive rate, 실제 음성구간인데 음성구간이라고 판단함)과 오검출율 (false positive rate, 실제 비음성구간인데 음성구간

으로 잘못 판단됨)이 모두 감소하게 된다. 문턱값은 음성구간 검출이 적용되는 시스템에 따라 다르게 결정되어야 하는데, 예를 들어 음성구간 검출이 잡음제거 시스템에 적용되어 잡음의 특성을 추정하는데 사용된다면 낮은 문턱값을 채택한다. 그 이유는 음성구간이 비음성구간으로 잘못 검출되는 경우를 최소화하여 잡음제거 성능을 낮추더라고 음성왜곡이 일어나지 않도록 하는 것이 중요하기 때문이다.

IV. 얼굴 및 입술 영역 검출

제안하는 영상 기반 음성 검출기는 입술 영역의 움직임과 픽셀 밝기 변화를 이용하기 때문에 얼굴 영역과 입술 영역을 검출하는 과정이 선행적으로 요구된다. 본 논문에서는 정면 얼굴 영상만을 고려하였는데, 보다 실용적인 알고리즘이 되기 위해서는 얼굴의 정면뿐만 아니라 측면도 이용하는 것[18]이 바람직하나 휴대폰과 같이 정면 얼굴을 계속적으로 취득할 수 있는 응용분야에 초점을 맞추고자 한다. 본 장에서는 본 논문에서 사용한 얼굴 및 입술 영역 검출 알고리즘에 대하여 설명한다.

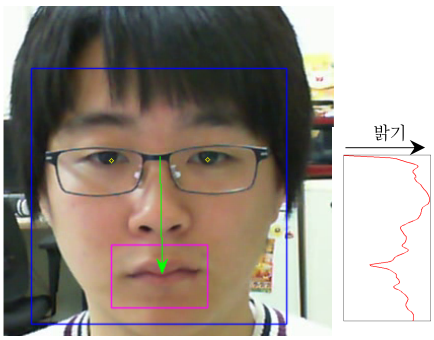
1. 얼굴 검출 알고리즘

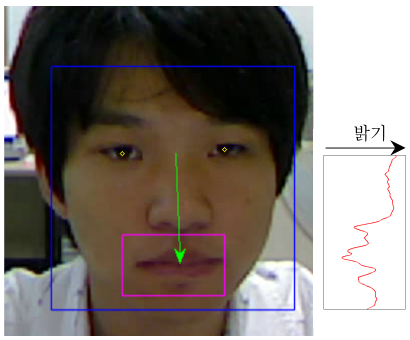
본 논문에서는 정면 얼굴을 검출하기 위하여, 가장 잘 알려져 있는 Viola, Jones의 알고리즘[19]을 사용한다. 이 알고리즘은 적분 영상을 이용하여 Haar-like 기저함수에 대한 특징 점을 빠르게 추출하고, 이렇게 얻어진 수많은 약분류기를 AdaBoost 알고리즘[20]으로 강화(Boosting)하여 강분류기(Strong Classifier)를 생성하여 학습한다. 그리고 좀 더 속도를 높이기 위해 검출 단계에서 위 과정을 통해 학습된 강분류기를 복잡도에 따라 계층적으로 적용한다. 즉, 상위 분류기에서 학습된 물체가 아니라고 검출될 경우 하위 분류기는 동작 하지 않고 넘어가기 때문에 그만큼 계산량이 감소되어 실시간으로 동작 가능하다.

제안하는 알고리즘에서는 얼굴 검출 알고리즘 구현을 위해, OpenCV에 포함되어있는 cvHaarDetectObjects 함수를 이용하였다.

2. 입술 영역 검출 알고리즘

얼굴 검출이 이루어진 뒤에 이를 바탕으로 입술 영역을 추정하게 된다. 기존 연구로는 색상 정보와 모양 정보를 이용하여 입술영역일 가능성이 높은 영역을 분류하는 방법[21,22]과 경계선 모델을 이용하여 입술과 피부의 경계선을 검출하는 방법[23] 등이 있고, 두 방법 모두 정확한 입술 검출을 목표로 한다. 그러나 본 논문에서 제안하는 알고리즘은 입술 주변의 영상을 이용하여 음성구간 검출을 수행하므로 입술을 정확하게 검출할 필요는 없다. 따라서 본 논문에서는 얼굴의 기하학적인 형태와 영상에서 닫힌 입술 사이의 밝기가 낮게 나타난다는 사실을 이용하여 입술 주변 영역을 검출하고, 이를 이용하여 음성구간 검출을 수행한다.





본 논문에서는 사람 얼굴에 대한 다음과 같은 관측 경험을 토대로 입술 영역을 추정한다. 입술 영역의 크기는 일반적으로 가로 길이가 두 눈동자 사이의 거리와 비슷하고, 세로 길이는 얼굴의 세로 길이에 비례한다(그림 3). 입술의 위치를 구하기 위해서는 얼굴 영상의 밝기 변화를 이용하는데, 사람이 발화하고 있지 않을 경우 닫힌 입술 사이에 그늘이 발생하여 입술 중앙 부분은 비교적 어두운 값을 가지게 된다. 또한 입술의 중앙 부분은 두 눈 사이를 이은 직선과 수직하며 두 눈 사이의 중앙을 지나는 직선위에 위치하게 된다.

먼저, 얼굴 검출을 위해 적용한 알고리즘에 눈 영역에 대한 Haar 분류기를 적용하여 두 눈을 검출한다. 그리고 다음과 같이 두 눈의 중앙점을 구한다.

$$(x_c, y_c) = \left(\frac{x_{le} + x_{re}}{2}, \frac{y_{le} + y_{re}}{2} \right) \tag{7}$$

여기서 (x_{le}, y_{le})는 영상에서 왼쪽 눈의 위치, (x_{re}, y_{re})는 오른쪽 눈의 위치이다. 두 눈의 중앙점을 시작점으로 하고 얼굴의 아래 부분을 향하는 벡터 중, 두 눈을 잇는 직선과 수직인 벡터를 $\vec{l}$이라 하고 아래 식과 같이 $\vec{l}$ 벡터를 따라 밝기를 검사하여 가장 낮은 밝기를 갖는 점을 닫힌 입술의 중앙점(x_{lc}, y_{lc})으로 추정한다.

$$(x_{lc}, y_{lc}) = \arg\min_{(x,y)\in \vec{l}} I(x,y) \tag{8}$$
$$\vec{l} = (x_c + a(k - y_c), k), \; k \subset [y_c, y_{end}]$$

여기서 $I(x,y)$는 (x,y)좌표에 해당하는 영상의 밝기이고, y_{fend}는 얼굴 영상 아래 경계에 해당하는 y축 좌표이며, a는 두 눈을 연결하는 직선의 기울기이다. 이렇게 구한 입술 중앙점 (x_{lc}, y_{lc})을 중심점으로 하여 입술 영역을 정하게 되는데, 앞에서 언급한 것과 같이 가로 길이는 일반적으로 두 눈동자 사이의 거리와 비슷하므로 눈 검출을 통해 얻어진 두 좌표 사이의 거리로 설정하고, 세로 길이는 얼굴의 세로 크기에 0.25를 곱하여 설정하였다. 그림 3에 검출된 얼굴 및 벡터 $\vec{l}$상에서의 픽셀 밝기 값의 변화와 일련의 과정을 통해 얻어진 입술 영역을 나타내었다.

IV. 실험 결과

제안한 알고리즘의 성능을 검증하기 위해 음성발화로 이루어진 동영상 25개에 대해 음성구간검출을 수행하였다. 동영상에서의 얼굴은 모두 정면을 보고 있으며, 각 동영상마다 1개 또는 2개의 문장이 발음되었다. 전체 프레임 수는 4356이고, 이 가운데 음성구간은 1563 프레임이며, 나머지 2793 프레임은 비음성구간이다. 영상은 초당 30프레임인

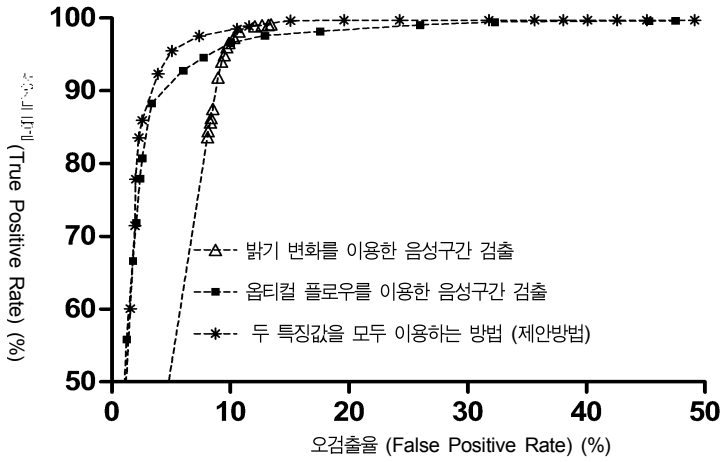


그림 4. 음성구간 검출 알고리즘의 ROC 곡선들
Fig. 4. ROC curves for voice activity detection

영상 10개와, 초당 15프레임인 영상 15개를 이용하였고, 음향 신호의 샘플링 주파수는 44.1kHz이다.

음성구간 검출 알고리즘의 성능 평가를 위해 여러 문턱값 (T)에 대해 정검출율과 오검출율을 계산하였다. 음성/비음성구간에 대한 참값은 음향파형으로부터 수작업을 통해 구하였다. 제안된 알고리즘의 성능 비교를 위해 옵티컬 플로우 기반의 특징값이나 밝기 변화 기반의 특징값만을 이용한 음성구간 검출 알고리즘의 성능도 평가하였다. 본 실험의 목적은 두 가지 특징값을 결합하는 것이 각각의 특징값을 사용하는 경우에 비해 우수하다는 것을 보이는 것이므로 서론에 열거한 다른 기존 알고리즘과의 비교는 수행하지 않았다. 그림 4에 문턱값 변화에 따른 정검출율과 오검출율을 ROC (Receiver Operating Characteristic) 곡선을 나타내었으며, 두 개의 특징값의 조합을 이용하는 제안된 알고리즘의 경우 약 5%의 오검출율을 갖도록 T값을 정했을 때 95% 정도의 정검출율을 보임을 알 수 있다. 본 실험결과에서 보듯이 하나의 특징값만을 이용한 알고리즘들에 비해 두 가지 특징값을 적절히 결합한 음성구간 검출기의 성능이 더 우수함을 알 수 있다.

V. 결 론

본 논문에서는 영상정보로부터 특징값을 추출하여 음성구간을 검출하는 알고리즘을 제안하였다. 영상정보를 이용하여 음성구간을 검출하는 방법은 음향잡음에 전혀 영향을 받지 않으므로 음향잡음이 심한 환경에서 음향신호를 이용하는 음성구간 검출을 대신할 수 있다. 제안된 알고리즘은 기존에 제안되었던 두 가지 영상정보 기반의 특징값을 기반으로 하고 있다. 첫 번째 특징값은 화자의 입술과 그 주변 영역의 움직임을 추정하는 옵티컬 플로우를 이용하는 것인데 본 논문에서는 기존의 방법과 달리 광역 움직임을 차감하여 얼굴전체의 움직임에 강인하도록 하였다. 두 번째 특징값은 발성시 드러나는 구강의 영향으로 어두운 화소가 증가한다는 사실에 기초하여 추출된다. 본 논문에서는 발성시 드러나는 구강의 어두운 화소뿐만 아니라 치아로 인한 밝은 화소도 고려하였다. 이와 같은 두 가지의 특징값을 추출한 뒤 각각의 특징값으로부터 스코어를 계산하여 하나의 최종 스코어를 구하고 이를 문턱값과 비교하여 음성구간을 검출하였다. 일반적으로 음성구간검출은 다른 알고리즘의 전처리로서의 기능을 하는 경우가 많으므로 적은 계산량이 요구된다. 이에 부합하기 위해 통계적 모델링과 같은 복잡한 방법 대신 문턱값 비교와 같은 비교적 간단한 계산량으로 음성구간을 추출할 수 있는 알고리즘을 제안하였다. 제안된 알고리즘의 성능을 검증하기 위하여 수행된 실험결과, 오검출율을 5%로 하였을 때 95% 정도의 검출율을 보이고 있으며 이는 하나의 특징값을 이용한 결과들와 비교하여 향상된 성능임을 확인하였다.

참 고 문 헌

[1] J. Sohn, N. S. Kim and W. Sung, “A statistical model-based voice activity detection,” IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1-3, January 1999.

[2] M. Hoffman, Z. Li, and D. Khataniar, “GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech,” IEEE Trans. on Speech and Audio Processing, vol. 9, no. 2, pp. 175-179, March 2001.

[3] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-27, no. 2, pp. 113-120, April, 1979.

[4] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, “An improved endpoint detector for isolated word recognition,”, IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-29, no. 4, pp. 777-785, August 1981.

[5] B.-F. Wu, “Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments,” IEEE Trans. Speech and Audio Processing, vol. 13, no. 5, pp. 762-775, September 2005.

[6] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, “Use of a CSP-based voice activity detector for distant-talking ASR,” in Proceedings of EUROSPEECH, Geneva, 2003.

[7] G. Kim and N. I. Cho, “Voice activity detection using phase vector in microphone array,” Electronics Letters, vol. 43, issue 14, pp. 783-784, July 2007.

[8] H. Yehia, R. Rubin, and E. Vatikiotis-Bateson, “Quantitative association of vocal-tract and facial behavior,” Speech Communication, vol. 26, no. 1, pp. 23-43, August 1998.

[9] P. Liu and Z. Wang, “Voice activity detection using visual information,” in Proceedings of ICASSP, pp. 609-612, Montreal, Canada, May 2004.

[10] T. Cootes, G. Edwards, and C. Taylor, “Active appearance models,”

IEEE trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, June 2001.

[11] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, L. Chambers, and C. Jutten, “Two novel visual voice activity detectors based on appearance models and retinal filtering,” in Proceedings of EUSIPCO, September 2007.

[12] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas, “Visual lip activity detection and speaker detection using mouth region intensities,” IEEE trans. Circuits and Systems for Video Technology, vol. 19, no. 1, pp. 133-137, January 2009.

[13] R. Navarathna, D. Dean, P. Lucey, S. Sridharan, and C. Fookes, “Cascading appearance-based features for visual voice activity detection,” in Proceedings of International Conference on Audio-Visual Speech Processing, Hakone, Japan, September 2010.

[14] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," Image Processing, IET, vol. 4, no. 4, pp. 463-472, December 2010.

[15] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," J. VLSI Signal Process. Syst., vol. 36, pp. 117-124, February 2004.

[16] D. Sun, S. Roth, and M. Black, “Secrets of optical flow estimation and their principles,” In Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2432-2439, San Francisco, USA, June 2010.

[17] B. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679, April 1981.

[18] R. Navarathna, D. Dean, S. Sridharan, C. Fookes, and P. Lucey, “Visual voice activity detection using frontal versus profile views,” In Proceedings of the International Conference on Digital Image Computing : Techniques and Applications, December 2011.

[19] P. Viola and M. Jones, “Robust Real-time Object Detection”, Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling, Vancouver, Canada, July 2001.

[20] Y. Freund and R.E. Schapire “A decision-theoretic generalization of on-line learning and an application to boosting”, In Computational Learning Theory: Eurocolt, Springer-Verlag, pp. 23 – 37, 1995

[21] E. Skodras and N. Fakotakis, “An Unconstrained Method for Lip Detection in Color Images”, in Proceedings of ICASSP, Prague, Czech, 2011.

[22] G. Fanelli and J. Gall and L. Van Gool, “Hough Transform-based Mouth Localization for Audio-Visual Speech Recognition”, British Machine Vision Conference, 2009

[23] X. Liu, Y. Cheung M. Li and H. Liu, “A Lip Contour Extraction Method Using Localized Active Contour Model with Automatic Parameter Selection”, 20th Int. Conf. on Pattern Recognition (ICPR), August 2010.

저 자 소 개

김 기 백

- 1994년 : 서울대학교 전자공학과 학사
- 1996년 : 서울대학교 전자공학과 석사
- 2007년 : 서울대학교 전기컴퓨터공학부 박사
- 1996년 ~ 2000년 : LG전자기술원 연구원
- 2000년 ~ 2003년 : (주)보이스웨어 선임연구원
- 2008년 ~ 2010년 : Univ. of Texas at Dallas, Research Associate
- 2010년 ~ 2011년 : 대구대학교 전자공학부 전임강사
- 2011년 ~ 현재 : 숭실대학교 조교수
- 주관심분야 : 음성신호처리, 영상신호처리, 멀티모달신호처리, 어레이신호처리

유 제 웅

- 2009년 2월 : 서울대학교 전기공학부 학사
- 2009년 3월 ~ 현재 : 서울대학교 전기컴퓨터공학부 박사과정
- 주관심분야 : 영상처리, 컴퓨터 비젼

저 자 소 개

조 남 익

- 1986년 : 서울대학교 제어계측공학과 학사
- 1988년 : 서울대학교 제어계측공학과 석사
- 1992년 : 서울대학교 제어계측공학과 박사
- 1991년 ~ 1994년 : 제어계측 신기술연구센터 연구원
- 1994년 ~ 1998년 : 서울시립대학교 전자전기공학부 조교수
- 1999년 ~ 현재 : 서울대학교 전기공학부, 조교수, 부교수, 교수
- 주관심분야 : 디지털 신호처리