구강 영역에 대한 타원 근사법을 이용한 음성 구간 검출법

류제웅, 추성권, *김기백, 조남익
서울대학교, *숭실대학교
youjw@ispl.snu.ac.kr, chewry@ispl.snu.ac.kr, *imkgb27@ssu.ac.kr, nicho@snu.ac.kr

Voice Activity Detection Using Ellipse Fitting of the Oral Cavity Region

Jewoong Ryu, Sung Kwon Choo, *Gibak Kim and Namik Cho
Seoul National University *Soongsil University

요 약

음성 신호처리에서 많이 쓰이는 음성구간 검출은 주로 음향신호의 분석을 통하여 음향 신호에 음성이 존재하는지 여부를 판단한다. 그러나 음향신호를 이용한 방법은 음성 또는 비음성 잡음이나 주위 음향 환경에 의하여 성능이 결정된다는 단점이 있다. 음향 환경 변화에 강인한 음성구간 검출을 수행하기 위하여, 영상정보를 이용한 음성구간 검출 방법들이 최근에 연구되어 왔는데 기존 방법들은 입술 모양의 변화를 추정하기 위하여 입술 모델 등을 이용하거나 구강(oral cavity) 영역에 해당하는 픽셀 수의 변화를 이용하여 음성 구간을 검출하였다. 위 방법들은 입술의 모양을 추정하는 데 복잡한 계산이 필요하거나, 입술 모양 추정 없이 구강 영역 픽셀 수만 이용하기 때문에 다소 정확도가 떨어진다는 단점이 있다. 본 논문에서는, 입술 모양의 변화를 추정하기 위해 밖으로 드러나는 구강 영역의 모양을 타원 근사법으로 추정하고, 타원의 넓이와 높이의 변화를 이용하여 음성 구간을 검출하는 방법을 제안하였다. 비교 실험 결과, 제안하는 방법은 구강영역 픽셀 수의 변화만 이용하는 방법에 비해 우수한 성능을 보임을 확인할 수 있었다.

1. 서론

음향 신호에서 음성의 포함 여부를 판별하는 음성 구간 검출 알고리즘은, 음성 인식, 잡음 제거, 음성 압축 등의 다양한 음성신호처리 분야에서 널리 사용되고 있다. 일반적으로 음성 구간 검출은 음향신호를 이용하여 수행되어 왔는데, 이 방법은 음향 잡음과 주위 음향 환경에 따라서 성능이 크게 달라지게 된다. 따라서 심한 음성 잡음이 존재하는 환경에서도 강인하게 음성 구간 검출을 수행하기 위하여, 마이크로폰 어레이를 사용하는 방법[1] 등이 연구되었으나, 이 방법의 경우에도 음향 잡음의 특성이 음성과 비슷하거나 신호 대 잡음 비(SNR)가 낮은 열악한 환경일 경우 성능이 떨어지게 된다.

최근 들어 영상신호와 음성신호를 동시에 입력 받을 수 있는 PC 카메라, 스마트 폰 등의 보급이 증가하였고, 영상 통화나 화상 회의 등 화자가 카메라를 정면으로 보고 있는 상황에서 영상 신호와 음성 신호를 동시에 처리하는 멀티모달 어플리케이션의 필요성의 증가하고 있다. 화자의 발화 영상정보를 이용하면 발화 과정에서 나타나는 입 주위 근육의 움직임이나 입술 모양의 변화와 같은 정보를 이용하여 발화 여부를 알아낼 수 있고[2], 이러한 영상 정보는 주위 음향 환경에 영향을 받지 않기 때문에 이를 바탕으로 음성 구간을 검출하는 알고리즘의 연구가 최근 진행되어 왔다[3-6].

영상정보를 이용한 음성 구간 검출 알고리즘 연구는 크게 입술 모델 추정을 이용하는 방법 [3,4]와 입술 영역의 밝기의 변화를 이용하는 방법[5,6]으로 나눌 수 있다. A. Aubrey 등이 제안한 방법은, Active Appearance Model(AAM)을 이용하여 입술 영역의 파라미터를 추출하여, 은닉 마르코프 모델(HMM)으로 이 값들을 훈련하여 음성 구간 검출을 수행하였다[3]. 또한, D. Sodoyer 등은, 입술 모양에 대한 파라미터를 Chroma Key 를 이용하여 추출한 뒤 모델 파라미터의 변화를 이용하여 음성구간 검출에 이용하였다[4]. 위 방법들은 입술 모양의 정확한 추정과 사전 훈련과정이 필요하다는 단점이 있다. 한편, S. Siatras 등은 발화 과정에서 드러나는 구강영역의 어두운 픽셀 수의 시간에 따른 증감을 통계적 모델로 근사하여 음성 구간 검출에 이용하였다[5]. 그리고 구강 영역 픽셀수의 증감과 옵티컬 플로우를 이용하여 추출된 주변 영역의 움직임을 특징값으로 사용한 방법도 제안되었다[6]. 위 두 방법은 비교적 계산량이 적고 구현이 간단하나, 구강 영역의 구조를 반영하지 않고 단순히 픽셀 수만 이용하므로, 영상 크기의 변화나 영상 노이즈 등에 의해 성능이 감소되는 단점이 있다.

본 논문에서는, 발화 시에 드러나는 구강영역의 변화를 타원 근사법을 이용한 모델링으로 추정하고 타원의 너비와 높이의 비율을 특징값으로 이용하여 음성구간을 판단하는 알고리즘을 제안한다. 제안하는 알고리즘은 구강영역의 모양을 타원으로 모델링하여 예측함으로써, 입술 모양의 변화를 [3,4]보다 비교적 간단하게 추적할 수 있을 뿐만 아니라, 타원 파라미터의 변화를 계산하여 음성 구간을 예측할 수 있다. 또한 기존에 제안한 알고리즘[6]과 비교 결과 오검출율 3%~5%

구간에서 91.8% ~ 93.23%의 검출율을 보여 기존 알고리즘보다 우수한 성능을 보이는 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2 장에서는 구강영역을 구분하고, 이를 타원 근사법을 통하여 모델링 하는 방법에 대해 기술하였다. 3 장에서는 타원 모델을 이용하여 특징값을 추출하고, 음성구간을 판정하는 방법에 대해 서술하였다. 그리고 4 장에서 기존 알고리즘 [6]과의 성능을 비교하였고, 5 장에 결론과 향후 과제를 논하였다.

2. 구강 영역 타원 근사법

2.1 입술 영역 검출

구강 영역을 찾기 위해서, 먼저 주어진 영상에서 입술 주변 영역을 지정하는 과정이 필요하다. 일반적으로 카메라를 통해 입력된 영상은 발화자의 얼굴과 그 이외의 영역으로 구성되어 있다. 이 영상에서 얼굴 영역을 분리하기 위해, 가장 널리 쓰이는 Viola-Jones 방법[7]을 이용하여 얼굴영역을 지정하고, 지정된 얼굴영역에서 역시 같은 방법으로 눈 영역을 지정한다. 이렇게 결정된 얼굴 영역과 눈의 위치를 바탕으로,

$$(x_{lc}, y_{lc}) = \arg\min_{(x,y)\in\vec{l}} I(x,y)$$
$$\vec{l} = (x_c + a(k - y_c), k), k \subset [y_c, y_{end}] \quad (1)$$

와 같이 입술의 중앙점을 찾을 수 있다. 여기서 (x_{lc}, y_{lc}) 는 구해진 입술의 중앙점이고, $I(x,y)$ 는 주어진 영상의 (x,y) 좌표에 해당하는 밝기이다. 그리고 (x_c, y_c) 는 두 눈 좌표의 중앙점이고, a 는 두 눈을 연결하는 직선의 기울기이다. 입술의 중앙점을 바탕으로 입술 영역을 설정하는데, 제안하는 알고리즘에서는 입술의 정확한 위치를 사용하지 않고, 입술을 포함하는 영역의 설정만 필요하므로, 단순하게 너비는 눈 사이의 거리로 설정하고 높이는 얼굴 높이의 1/4 으로 설정한다. 이렇게 구해진 입술 영역을 그림 1 에 나타내었다.

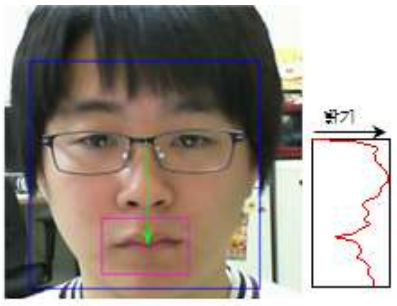

그림 1 입술 영역 검출 과정 및 결과. 녹색 벡터는 $\vec{l}$ 을 나타내고, 오른쪽 그래프는 영상에서 $\vec{l}$ 에 해당하는 픽셀 값의 변화를 나타낸다.

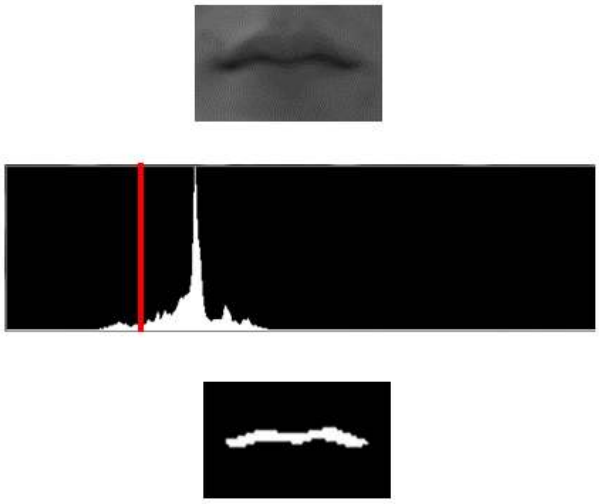

그림 2 (위) 입술 영역 영상.
(중간) 입술 영역 영상의 히스토그램과 문턱 값.
(아래) 구강 영역으로 설정된 이진 영상

2.2 구강 영역 설정

2.1 절에서 구한 입술 영역을 바탕으로, 구강 영역을 예측한다. 구강 영역은 입을 다물고 있을 때는 드러나지 않다가, 발화가 시작되어 입을 움직이면 드러나게 된다. 그런데, 구강은 입 안쪽에 위치하므로 빛이 잘 닿지 않기 때문에 일반적으로 주변의 피부나 입술보다 어두운 밝기를 가지게 된다. 이런 관측을 바탕으로, 화자가 발화하지 않는 초기 상태일 때, 그림 2 와 같이 입술 영역의 히스토그램을 계산한 뒤, 하위 5%의 값을 문턱값으로 설정한다. 문턱값을 τ 라 하면,

$$I_{cav}(x,y) = \begin{cases} 1 & I(x,y) < \tau \\ 0 & otherwise \end{cases} \quad (2)$$

를 이용하여 구강 영역에 해당하는 이진 영상을 얻을 수 있다.

2.3 타원 근사법을 통한 구강영역 예측

위에서 구한 구강 영역에 해당하는 이진 영상을 이용하여, 타원 모델을 이용하여 구강의 모양을 근사하기 위해, 이진영상 중 가장자리에 해당하는 몇 개의 픽셀만 이용하면 정확도와 효율을 높일 수 있다. 그러므로, 먼저 이진영상에서 가장자리에 해당하는 픽셀들을 추출한 뒤 균일하게 30%에 해당하는 점들만 뽑고, 이 점들을 이용하여 타원 근사법을 통해 구강 영역을 예측한다. 본 논문에서는 [8]에서 제안한 방법을 이용하여 구강 영역의 가장자리 점들을 추출하였다. 추출한 가장자리 점들을 그림 3 에 나타내었다.

그림 3 추출된 구강 영역의 가장자리 점들

이렇게 구한 가장 자리 점들을 이용하여 타원 근사법을 수행하는데, 본 논문에서는 최소자승법을 이용한 타원 근사법을 이용하였다[9]. 주요 모음별 입술 영역 영상과 위 과정을 거쳐 구분된 구강 영역 이진 영상 및 추출된 가장자리 값과 타원 근사법을 통하여 예측된 타원 모양을 그림 4 에 나타내었다

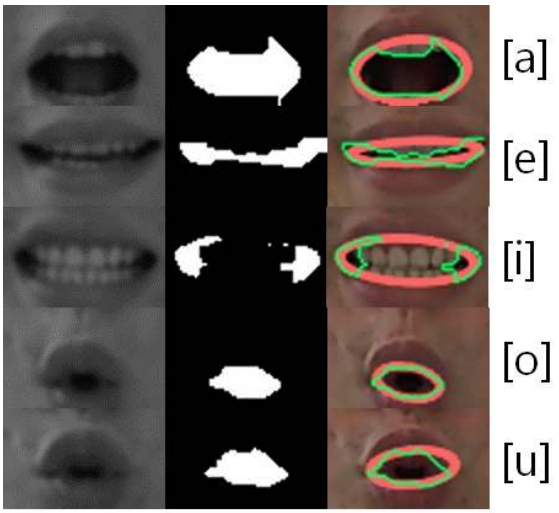

그림 4 주요 모음 발음에 대한 입술 영역 영상, 구강영상 및 가장자리 추출과 타원 근사 결과. 주황색 선이 타원 근사를 수행한 결과이고, 녹색 선이 추출된 가장자리 픽셀이다.

그림 4 와 같이, [a], [o], [u]발음 같이 이가 드러나지 않는 경우뿐만 아니라, [e]와 [i]발음 같이 이가 드러나서 구강영역의 이진 영상이 정확하게 구해지지 않은 경우에도, 타원 근사법을 통하여 구강영역의 모양을 비교적 정확하게 예측하는 것을 알 수 있다.

3. 음성 구간 검출

3.1. 특징값 추출

이번 절에서는 음성 구간 검출에 사용하기 위한 특징값을 추출하는 방법에 대해 설명한다. 사람이 발화를 시작하게 되면 입이 움직이게 되어 구강영역이 드러나면서 추정한 타원의 모양이 변화하게 되는데, 이 변화를 특징값으로 나타내어 음성 구간 여부를 판단하게 된다. 그림 4 에도 나타나듯이, 발화 시에는 그렇지 않을 때보다 예측된 구강영역의 높이가 크게 나타나고, 너비는 비슷하거나 줄어드는 경향을 보이게 된다. 따라서 입의 움직임에 따른 예측된 타원의 크기 변화를 충분히 반영할 수 있도록,

$$r_n = \frac{h_n / w_n}{r_{sil}} + \left| \frac{h_n / w_n}{r_{sil}} - r_{n-1} \right| \qquad (3)$$

$$\text{where, } r_{sil} = \sum_{i=s}^{s+N-1} \frac{h_n / w_n}{N}$$

을 특징값으로 사용한다. 여기서 h_n 과 w_n 은 각각 n 번째 프레임의 예측된 타원의 높이와 너비를 나타낸다. 그리고 r_{sil} 은 비발화 구간의 타원 너비에 대한 높이의 비율을 나타내는 값으로, 묵음 구간인 s 번째 프레임부터 N 프레임 동안의 타원의 너비에 대한 높이의 비율을 평균 내어 얻은 값이다. 또한 $\left| \frac{h_n / w_n}{r_{sil}} - r_{n-1} \right|$ 항을 추가하여, 일시적으로 현재 프레임의 $\frac{h_n / w_n}{r_{sil}}$ 이 작은 값을 가지더라도 큰 특징값을 가지도록 하여 검출율을 높일 수 있도록 한다. 그림 5 에 한 문장에 대하여 실제 음성 구간과 제안하는 특징값의 변화를 도시하였다. 제안한 특징 값이 비교적 음성 구간 여부를 잘 나타내는 것을 볼 수 있다.

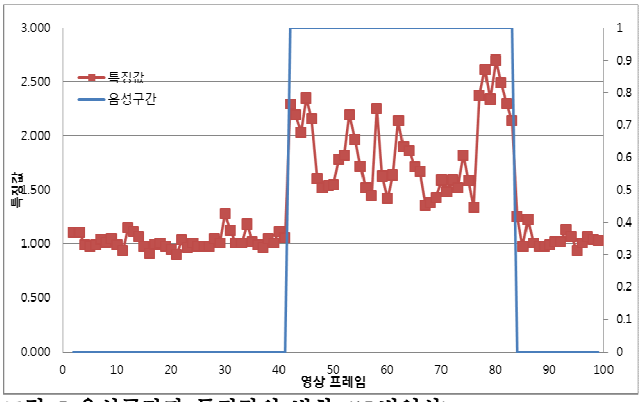

그림 5 음성구간과 특징값의 변화 (15 번영상)

3.2. 음성 구간 판단

음성 구간 판단을 위하여 추출한 특징값은, 그림 5 에서도 나타나듯이 음성 구간 내에서도 증감을 반복하여 한 개의 문턱값으로 판별하게 되면 성능이 떨어지게 될 위험성이 있다. 따라서 본 논문에서는,

$$score_{r_n} = \begin{cases} 1.0 & , r_n > \tau_{voice} \\ \dfrac{1.0}{\tau_{voice} - \tau_{sil}} (r_n - \tau_{sil}) & , \tau_{sil} < r_n \le \tau_{voice} \\ 0.0 & , r_n \le \tau_{sil} \end{cases} \qquad (4)$$

와 같이 두 개의 문턱 값을 사용하고, 그 사이의 값은 선형적으로 증가하도록 점수를 산정하여 위와 같은 상황에 대응할 수 있도록 하였다. 여기서 r_n 3.1 에서 추출한 특징값이고, τ_{voice} 와 τ_{sil} 는 각각 음성구간과 비음성 구간을 나타내는 문턱 값으로 본 논문의 실험에서는 각각 1.05 와 1.7 을 사용하였다. 마지막으로 음성 구간 판정은,

$$VAD = \begin{cases} \text{TRUE}, & score_{r_n} > T \\ \text{FALSE}, & score_{r_n} \le T \end{cases} \qquad (5)$$

와 같이 판정을 내리게 된다. 여기서 T 는 검출기의 민감도를 결정하는 상수로, 0 과 1 사이의 값을 가지며 0 에 가까울수록 검출율과 오검출율이 증가하고, 1 에 가까울수록 검출율과 오검출율이 감소하게 된다.

4. 실험 결과

제안한 음성 검출 알고리즘의 성능을 평가하기 위해, 총 27 개의 영상을 대상으로 실험을 수행하였다. 각 영상은 두 문장 이하의 발화를 포함하고 있고, 총 4842 프레임 중 비발화 프레임이 2951, 발화 부분에 해당하는 프레임이 1891 이었다. 성능 비교를 위하여, [6]에서 제안한 구강 영역의 어두운 픽셀 수를 이용한 방법과, 어두운 픽셀과 움직임을 동시에 고려한 방법의 실험도 같은 데이터에 대해 수행하였다. 각 알고리즘에 대하여 검출율(TPR)과 오검출율(FPR)을 구하여, 수신자 조작 특성 곡선(ROC Curve)를 그림 6 에 도시하였다.

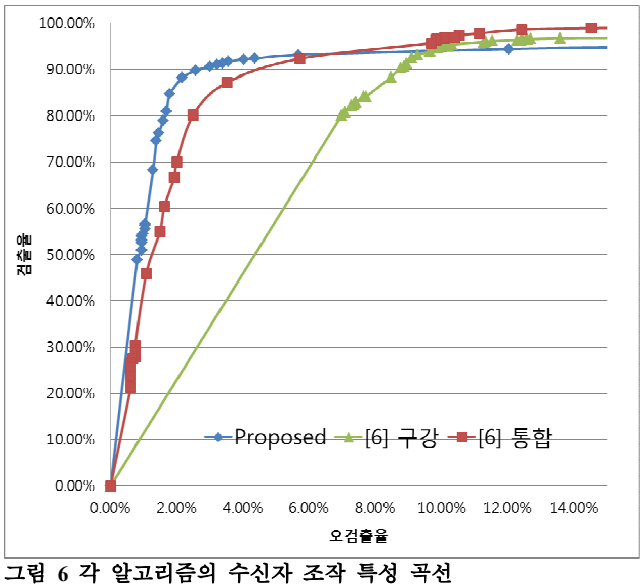

그림 6 각 알고리즘의 수신자 조작 특성 곡선

실험 결과에 나타나듯이, 제안하는 방법이 기존에 제안된 구강 영역의 픽셀 수만 고려하는 방법([6] 구강)과 움직임 정보도 같이 고려하는 방법([6] 통합)에 비해, 오검출율 2~6%구간에서 약 88%~94%의 검출율을 보여 가장 좋은 성능을 나타낸다는 것을 알 수 있다. 제안하는 알고리즘은 구강 영역의 타원 모델 추정 과정을 바탕으로 비교적 정확하게 드러난 구강 영역의 크기를 나타낼 수 있기 때문에, 이가 드러나는 경우나 영상 잡음에 의하여 구강 영역 픽셀 수를 정확하게 계산하지 못하는 [6]에 비해 좋은 성능을 나타내었다. 또한 [6]에서 제안한 구강영역의 픽셀 수 변화와 움직임을 동시에 고려하는 방법에 비해서도 우수한 성능을 나타냄을 확인할 수 있다.

5. 결론 및 향후 과제

본 논문에서는 구강 영역을 타원 근사법을 이용하여 예측하고, 타원 파라미터를 이용하여 음성 구간을 검출하는 방법을 제안하였다. 제안하는 방법은 사람이 발화 시에 드러나는 구강 영역은 주변부보다 어둡다는 사실을 이용하여 구강 영역의 이진 영상과 가장자리 영역을 추출한 뒤, 타원 근사법을 이용하여 추정하였다. 이 방법을 통하여 발화 시 입의 모양을 비교적 정확하게 예측하였으며, 타원의 장축과 단축 비율의 변화를 특징값으로 음성 구간 검출에 적용하여 기존에 단순히 구강 영역의 픽셀 수만 고려하는 알고리즘보다 좋은 성능을 나타내었다. 본 논문에서 제안한 타원 모델은 적은 계산량으로 입 모양 예측이 가능하므로, 향후 연구에서는 이 모델을 이용하여 더 향상된 성능의 음성 구간 검출이나 다른 응용 분야에 적용 등을 수행할 예정이다.

감사의 글

이 논문은 2011 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행되었습니다 (2012-0003455). 또한 이 논문은 지식경제부 지원으로 수행하는 21 세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행되었습니다.

6. 참고문헌

[1] 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.

[2] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, " Quantitative association of vocal-tract and facial behavior" , Speech Communication, vol. 26, no. 1, pp. 23-43, 1998.

[3] 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.

[4] D. Sodoyer, B. Rivet, L. Girin, J.L. Schwartz and C. Jutten, " An Analysis of Visual Speech Information Applied to Voice Activity Detection," in Proc. ICASSP, 2006

[5] S. Siatras, N. Nikolaidis, M. Krinidis and I. Pitas, "Visual Lip Activity Detection and Speaker Detection Using Mouth Region Intensities", IEEE Trans. on Circuit Systems for Video Technology, Vol 19, No. 1, Jan. 2009

[6] 류제웅, 김세윤, 하성종, 조남익, " 입술 주변의 움직임과 밝기정보를 이용한 음성 구간 검출방법" , 제 24 회 신호처리통합학술대회, 제 24 권 1 호, 2011 년 9 월

[7] 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.

[8] S. Suzuki and K. Abe, " Topological structural analysis of digitized binary images by border following" , Computer Vision, Graphics and Image Processing. Vol. 30, Issue 1, pp. 32-46, April 1985

[9] A. Fitzgibbon, M. Pilu and R. B. Fisher, " Direct Least Square Fitting of Ellipses" , IEEE Trans. On Pattern Analysis and Machine Inteligence(PAMI), Vol. 21, No. 5 pp. 476-480, May 1999