



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0129676
(43) 공개일자 2017년11월27일

- | | |
|--|---|
| <p>(51) 국제특허분류(Int. Cl.)
 <i>G10L 25/93</i> (2013.01) <i>G06T 7/20</i> (2017.01)
 <i>G10L 15/25</i> (2013.01) <i>G10L 25/78</i> (2013.01)</p> <p>(52) CPC특허분류
 <i>G10L 25/93</i> (2013.01)
 <i>G06T 7/20</i> (2013.01)</p> <p>(21) 출원번호 10-2017-7015391</p> <p>(22) 출원일자(국제) 2015년03월19일
 심사청구일자 없음</p> <p>(85) 번역문제출일자 2017년06월05일</p> <p>(86) 국제출원번호 PCT/KR2015/002654</p> <p>(87) 국제공개번호 WO 2016/148322
 국제공개일자 2016년09월22일</p> | <p>(71) 출원인
 삼성전자주식회사
 경기도 수원시 영통구 삼성로 129 (매탄동)
 고려대학교 산학협력단
 서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)</p> <p>(72) 발명자
 고한석
 서울시 용산구 이촌로88길 30 103동 1602호
 김성수
 경기도 부천시 소사구 은성로 4
 (뒷면에 계속)</p> <p>(74) 대리인
 리앤목특허법인</p> |
|--|---|

전체 청구항 수 : 총 15 항

(54) 발명의 명칭 영상 정보를 기반으로 음성 구간을 검출하는 방법 및 장치

(57) 요약

적어도 하나의 영상에서 입술이 존재하는 영역을 검출하고, 상기 검출된 영역에 포함된 픽셀값들의 차이에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 획득하여, 상기 특징값에 기초하여, 상기 적어도 하나의 영상으로부터 상기 음성 구간을 검출하는, 방법이 개시된다.

(52) CPC특허분류

G10L 15/25 (2013.01)

G10L 25/78 (2013.01)

G10L 2025/783 (2013.01)

(72) 발명자

송태엽

서울시 노원구 동일로228길 23 1403동 701호

이경선

인천시 남구 인하로 230-34 301호

이재원

서울시 강남구 선릉로126길 22 102동 1001호

명세서

청구범위

청구항 1

음성 구간을 검출하는 방법에 있어서,

적어도 하나의 영상에서 입술이 존재하는 영역을 검출하는 단계;

상기 검출된 영역에 포함된 픽셀값 들간 차이에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 획득하는 단계;

상기 특징값에 기초하여, 상기 적어도 하나의 영상으로부터 상기 음성 구간을 검출하는 단계를 포함하는, 방법.

청구항 2

제1항에 있어서, 상기 특징값을 획득하는 단계는

기준 픽셀 및 상기 기준 픽셀의 주변 픽셀들의 평균값을 획득하는 단계;

상기 평균값과 상기 기준 픽셀 및 주변 픽셀 간 차이값에 기초하여 상기 특징 값을 획득하는 단계를 포함하는, 방법.

청구항 3

제1항에 있어서, 상기 특징값을 획득하는 단계는

상기 각 픽셀 및 주변 픽셀들에 대한 대표값과 상기 주변 픽셀 간 차이에 기초하여, 상기 각 픽셀의 분산 값을 획득하는 단계;

상기 각 픽셀의 분산값에 기초하여, 상기 적어도 하나의 영상의 특징값을 획득하는 단계를 포함하는, 방법.

청구항 4

제1항에 있어서, 상기 특징값을 획득하는 단계는

상기 차이가 미리 설정된 임계값보다 큰 픽셀의 개수를 상기 적어도 하나의 영상에 대하여 획득하는 단계;

상기 영상의 이전 영상 또는 이후 영상 간에 상기 획득된 픽셀 수의 차이를 상기 영상의 특징값으로 획득하는 단계를 포함하는, 방법.

청구항 5

제1항에 있어서, 상기 음성 구간을 검출하는 단계는

상기 특징값이 제1 임계값보다 커지면, 상기 특징값의 시점을 음성 구간의 시작점으로 결정하는 단계;

상기 특징값이 상기 제1 임계값보다 작아지면, 카운트 값을 0으로 설정하는 단계;

상기 특징값이 상기 제1 임계값보다 작아진 시점부터 상기 특징값이 상기 제1 임계값 및 제2 임계값 사이에 존재하는 동안, 상기 카운트 값을 시간이 경과함에 따라 증가시키는 단계;

상기 카운트 값이 미리 설정된 갭(gap) 크기보다 커진 시점을 상기 음성 구간의 종료점으로 결정하는 단계를 포함하는, 방법.

청구항 6

음성 구간을 검출하는 방법에 있어서,

적어도 하나의 영상에서 입술이 존재하는 영역을 검출하는 단계;

상기 검출된 영역에 포함된 각 픽셀에 대한 움직임 정보를 획득하는 단계;

분할된 영역이 서로 대칭되도록 상기 검출된 영역을 분할하는 단계;

상기 분할된 영역의 움직임 정보의 차이값에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 상기 적어도 하나의 영상에 대하여 획득하는 단계;

상기 적어도 하나의 영상의 특징값에 기초하여, 상기 음성 구간을 검출하는 단계를 포함하는, 방법.

청구항 7

제6항에 있어서, 상기 움직임 정보를 획득하는 단계는

옵티컬 플로우 방법을 이용하여 상기 각 픽셀에 대한 움직임 정보를 획득하는 단계를 포함하는, 방법.

청구항 8

음성 구간을 검출하는 디바이스에 있어서,

사용자의 얼굴을 포함하는 적어도 하나의 영상을 수신하는 수신부;

상기 적어도 하나의 영상에서 입술이 존재하는 영역을 검출하고, 상기 검출된 영역에 포함된 픽셀 값들간 차이에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 획득하고, 상기 특징값에 기초하여, 상기 적어도 하나의 영상으로부터 상기 음성 구간을 검출하고, 상기 검출된 음성 구간과 대응되는 음향 신호에 기초하여, 음성 인식을 수행하는 제어부를 포함하는 디바이스.

청구항 9

제8항에 있어서, 상기 제어부는

기준 픽셀 및 상기 기준 픽셀의 주변 픽셀들의 평균값을 획득하고, 상기 평균값과 상기 기준 픽셀 및 주변 픽셀 간 차이값에 기초하여 상기 특징 값을 획득하는, 디바이스.

청구항 10

제8항에 있어서, 상기 제어부는

상기 각 픽셀 및 주변 픽셀들에 대한 대표값과 상기 주변 픽셀 간 차이에 기초하여, 상기 각 픽셀의 분산 값을 획득하고, 상기 각 픽셀의 분산값에 기초하여, 상기 적어도 하나의 영상의 특징값을 획득하는, 디바이스.

청구항 11

제8항에 있어서, 상기 제어부는

상기 각 픽셀 및 주변 픽셀들에 대한 대표값과 상기 주변 픽셀 간 차이에 기초하여, 상기 각 픽셀의 분산 값을 획득하고, 상기 각 픽셀의 분산값에 기초하여, 상기 적어도 하나의 영상의 특징값을 획득하는, 디바이스.

청구항 12

제8항에 있어서, 상기 제어부는

상기 특징값이 제1 임계값보다 커지면, 상기 특징값의 시점을 음성 구간의 시작점으로 결정하고, 상기 특징값이 상기 제1 임계값보다 작아지면, 카운트 값을 0으로 설정하고, 상기 특징값이 상기 제1 임계값보다 작아진 시점부터 상기 특징값이 상기 제1 임계값 및 제2 임계값 사이에 존재하는 동안, 상기 카운트 값을 시간이 경과함에 따라 증가시키고, 상기 카운트 값이 미리 설정된 갭(gap) 크기보다 커진 시점을 상기 음성 구간의 종료점으로 결정하는, 디바이스.

청구항 13

음성 구간을 검출하는 디바이스에 있어서,

사용자의 얼굴을 포함하는 적어도 하나의 영상을 수신하는 수신부;

상기 적어도 하나의 영상에서 입술이 존재하는 영역을 검출하고, 상기 검출된 영역에 포함된 각 픽셀에 대한 움직임 정보를 획득하고, 분할된 영역이 서로 대칭되도록 상기 검출된 영역을 분할하고, 상기 분할된 영역의 움직임 정보의 차이값에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 상기 적어도 하나의 영상에 대하여 획득하고, 상기 적어도 하나의 영상의 특징값에 기초하여, 상기 음성 구간을 검출하고, 상기 검출된 음성

구간과 대응되는 음향 신호에 기초하여 음성 인식을 수행하는 제어부를 포함하는 디바이스.

청구항 14

제13항에 있어서, 상기 제어부는

옵티컬 플로우 방법을 이용하여 상기 각 픽셀에 대한 움직임 정보를 획득하는, 디바이스.

청구항 15

제1항 내지 제7항 중 어느 한 항에 있어서, 상기 방법을 구현하기 위한 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체.

발명의 설명

기술 분야

[0001] 본 발명은 입술 영역의 움직임 정보에 기초하여 음성 구간을 검출하는 방법 및 장치에 대한 것이다.

배경 기술

[0002] 음성 구간 검출은 음성 인식을 수행하기 위하여, 입력된 음향 신호로부터 음성 구간과 비음성 구간을 구별하는 것이다. 디바이스에서 음성 인식을 수행할 때, 음성 구간만의 음향 신호를 취하는 경우, 음성 인식에 걸리는 시간을 단축할 수 있다. 그러나 잡음이 심한 환경, 예를 들면 차량 주행 중인 상황에서는 다양한 잡음이 존재하므로, 음향 신호만을 이용하여 음성 구간을 검출하는 것은 한계가 존재한다.

[0003] 영상 신호를 이용하여, 음성 구간을 검출하는 기존 방법들은 대부분 실내 환경에서 촬영된 영상의 특징을 모델링하는 단계를 포함한다. 그러나 차량 환경에서는 실내와는 달리 조명이 계속 변화하므로 기존 방법을 이용하여 음성 구간을 정확히 검출하는 것은 어렵다.

발명의 내용

과제의 해결 수단

[0004] 본 발명은 입술 영역의 특징에 기초하여 음성 구간을 검출하는 방법 및 장치에 대한 것이다. 구체적으로, 입술 영역에서의 특징값의 변화에 기초하여 음성 구간을 검출하는 방법 및 장치에 대한 것이다.

발명의 효과

[0005] 일 실시 예에 의하면, 실내 조명 변화 및 얼굴 움직임과 같은 환경 변화에 강인한 입술 움직임의 특징값에 기초하여 영상에 기반한 음성 구간을 검출할 수 있다.

도면의 간단한 설명

[0006] 도 1은 일 실시 예에 의한 입술 영역의 특징에 기초하여 음성 구간을 검출하는 디바이스를 나타낸 블록도이다.

도 2는 일 실시 예에 있어서, 입술 영역을 검출하는 일 예를 도시한 예시도이다.

도 3은 일 실시 예에 의한 $VAR_{P,R}$ 값을 구하는 방법의 일 예를 나타내는 예시 도면이다.

도 4는 일 실시 예에 의한 LVH의 일 예를 나타낸 예시 도면이다.

도 5는 일 실시 예에 의한 $x(n)$ 값을 도시한 그래프를 나타낸 도면이다.

도 6은 일 실시 예에 의한 에지 검출 필터의 일 예를 그래프로 나타낸 예시 도면이다.

도 7은 일 실시 예에 의한 상태 천이 모델의 일 예를 나타낸 예시 도면이다.

도 8은 일 실시 예에 의한 $x(n)$, $F(n)$ 및 음성 구간으로 검출된 구간의 일 예를 나타낸 그래프이다.

도 9는 일 실시 예에 의한 묵음 구간의 일 예를 나타낸 예시 도면이다.

도 10은 일 실시 예에 의한 음성 구간을 검출하는 방법을 나타낸 순서도이다.

도 11은 일 실시 예에 의한 입술 움직임에 기초하여 음성 구간을 검출하는 방법을 나타낸 순서도이다.

도 12는 일 실시 예에 의한 입술 영역을 분할하는 일 예를 나타낸 예시 도면이다.

도 13은 일 실시 예에 의한 입술 움직임에 기초하여 음성 구간을 검출하는 디바이스의 내부 구조를 나타낸 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0007] 발명의 실시를 위한 최선의 형태

[0008] 일 실시 예에 의한 음성 구간을 검출하는 방법에 있어서, 적어도 하나의 영상에서 입술이 존재하는 영역을 검출하는 단계; 상기 검출된 영역에 포함된 픽셀값 들간 차이에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 획득하는 단계; 상기 특징값에 기초하여, 상기 적어도 하나의 영상으로부터 상기 음성 구간을 검출하는 단계를 포함한다.

[0009] 더하여, 상기 특징값을 획득하는 단계는 기준 픽셀 및 상기 기준 픽셀의 주변 픽셀들의 평균값을 획득하는 단계; 상기 평균값과 상기 기준 픽셀 및 주변 픽셀 간 차이값에 기초하여 상기 특징 값을 획득한다.

[0010] 더하여, 상기 특징값을 획득하는 단계는 상기 각 픽셀 및 주변 픽셀들에 대한 대표값과 상기 주변 픽셀 간 차이에 기초하여, 상기 각 픽셀의 분산 값을 획득하는 단계; 상기 각 픽셀의 분산값에 기초하여, 상기 적어도 하나의 영상의 특징값을 획득하는 단계를 포함한다.

[0011] 더하여, 상기 특징값을 획득하는 단계는 상기 차이가 미리 설정된 임계값보다 큰 픽셀의 개수를 상기 적어도 하나의 영상에 대하여 획득하는 단계; 상기 영상의 이전 영상 또는 이후 영상 간에 상기 획득된 픽셀 수의 차이를 상기 영상의 특징값으로 획득한다.

[0012] 더하여, 상기 음성 구간을 검출하는 단계는 상기 특징값이 제1 임계값보다 커지면, 상기 특징값의 시점을 음성 구간의 시작점으로 결정하는 단계; 상기 특징값이 상기 제1 임계값보다 작아지면, 카운트 값을 0으로 설정하는 단계; 상기 특징값이 상기 제1 임계값보다 작아진 시점부터 상기 특징값이 상기 제1 임계값 및 제2 임계값 사이에 존재하는 동안, 상기 카운트 값을 시간이 경과함에 따라 증가시키는 단계; 상기 카운트 값이 미리 설정된 갭(gap) 크기보다 커진 시점을 상기 음성 구간의 종료점으로 결정한다.

[0013] 일 실시 예에 의한 음성 구간을 검출하는 방법에 있어서, 적어도 하나의 영상에서 입술이 존재하는 영역을 검출하는 단계; 상기 검출된 영역에 포함된 각 픽셀에 대한 움직임 정보를 획득하는 단계; 분할된 영역이 서로 대칭되도록 상기 검출된 영역을 분할하는 단계; 상기 분할된 영역의 움직임 정보의 차이값에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 상기 적어도 하나의 영상에 대하여 획득하는 단계; 상기 적어도 하나의 영상의 특징값에 기초하여, 상기 음성 구간을 검출하는 단계를 포함한다.

[0014] 더하여, 상기 움직임 정보를 획득하는 단계는 옵티컬 플로우 방법을 이용하여 상기 각 픽셀에 대한 움직임 정보를 획득하는 단계를 포함한다.

[0015] 일 실시 예에 의한 음성 구간을 검출하는 디바이스에 있어서, 사용자의 얼굴을 포함하는 적어도 하나의 영상을 수신하는 수신부; 상기 적어도 하나의 영상에서 입술이 존재하는 영역을 검출하고, 상기 검출된 영역에 포함된 픽셀 값들간 차이에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 획득하고, 상기 특징값에 기초하여, 상기 적어도 하나의 영상으로부터 상기 음성 구간을 검출하고, 상기 검출된 음성 구간과 대응되는 음향 신호에 기초하여, 음성 인식을 수행하는 제어부를 포함한다.

[0016] 일 실시 예에 의한 음성 구간을 검출하는 디바이스에 있어서, 사용자의 얼굴을 포함하는 적어도 하나의 영상을 수신하는 수신부; 상기 적어도 하나의 영상에서 입술이 존재하는 영역을 검출하고, 상기 검출된 영역에 포함된 각 픽셀에 대한 움직임 정보를 획득하고, 분할된 영역이 서로 대칭되도록 상기 검출된 영역을 분할하고, 상기 분할된 영역의 움직임 정보의 차이값에 기초하여, 상기 검출된 영역의 입술 움직임에 대한 특징값을 상기 적어도 하나의 영상에 대하여 획득하고, 상기 적어도 하나의 영상의 특징값에 기초하여, 상기 음성 구간을 검출하고, 상기 검출된 음성 구간과 대응되는 음향 신호에 기초하여 음성 인식을 수행하는 제어부를 포함한다.

[0017] 발명의 실시를 위한 형태

[0018] 이하 본 발명의 바람직한 실시 예를 첨부한 도면을 참조하여 상세히 설명한다. 다만, 하기의 설명 및 첨부된 도면에서 본 발명의 요지를 흐릴 수 있는 공지 기능 또는 구성에 대한 상세한 설명은 생략한다. 또한, 도면 전체

에 걸쳐 동일한 구성 요소들은 가능한 한 동일한 도면 부호로 나타내고 있음에 유의하여야 한다.

- [0019] 이하에서 설명되는 본 명세서 및 청구범위에 사용된 용어나 단어는 통상적이거나 사전적인 의미로 한정해서 해석되어서는 아니 되며, 발명자는 그 자신의 발명을 가장 최선의 방법으로 설명하기 위한 용어로 적절하게 정의할 수 있다는 원칙에 입각하여 본 발명의 기술적 사상에 부합하는 의미와 개념으로 해석되어야만 한다. 따라서 본 명세서에 기재된 실시 예와 도면에 도시된 구성은 본 발명의 가장 바람직한 일 실시 예에 불과할 뿐이고, 본 발명의 기술적 사상을 모두 대변하는 것은 아니므로, 본 출원시점에 있어서 이들을 대체할 수 있는 다양한 균등물과 변형 예들이 있을 수 있음을 이해하여야 한다.
- [0020] 첨부 도면에 있어서 일부 구성요소는 과장되거나 생략되거나 또는 개략적으로 도시되었으며, 각 구성요소의 크기는 실제 크기를 전적으로 반영하는 것이 아니다. 본 발명은 첨부한 도면에 그려진 상대적인 크기나 간격에 의해 제한되어지지 않는다.
- [0021] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에서 사용되는 "부"라는 용어는 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소를 의미하며, "부"는 어떤 역할들을 수행한다. 그렇지만 "부"는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부"는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 "부"는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 "부"들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 "부"들로 결합되거나 추가적인 구성요소들과 "부"들로 더 분리될 수 있다.
- [0022] 아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0023] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예를 설명한다.
- [0024] 도 1은 일 실시 예에 의한 입술 영역의 특징에 기초하여 음성 구간을 검출하는 디바이스를 나타낸 블록도이다.
- [0025] 일 실시 예에 의한 디바이스(100)는 음향 신호와 대응되는 영상 신호에 포함된 사용자의 입술 영역의 특징을 분석함으로써, 음향 신호의 음성 구간을 검출할 수 있다. 디바이스(100)가 음성 인식을 수행할 때, 영상 분석을 통해 사용자가 발화 중인 음성 구간을 검출할 수 있다. 디바이스(100)는 영상에 기초하여 음성 구간을 검출할 수 있으나, 더하여 대응되는 음향 신호를 더 고려하여 음성 구간을 검출할 수도 있다. 디바이스(100)는 검출된 음성 구간과 대응되는 음향 신호에 대하여 음성 인식을 수행하면 높은 정확도로 음성 인식을 수행할 수 있다.
- [0026] 디바이스(100)는 실시간으로 사용자의 얼굴이 촬영된 영상의 입술 영역의 특징을 분석하여, 음성 구간을 검출할 수 있다. 발화 상황에서의 입술의 움직임 및 형태는 발화 전 상황에서의 입술의 움직임 및 형태와 차이가 있다. 예를 들면, 입술이 움직이는 경우, 발화 상황인 것으로 판단될 수 있다. 따라서, 디바이스(100)는 영상에서 입술 움직임 및 형태의 특징을 분석함으로써, 사용자가 발화 중인 구간으로 판단되는 음성 구간을 검출할 수 있다.
- [0027] 도 1을 참조하면, 디바이스(100)는 입술 영역 검출부(110), 입술 영역 움직임 분석부(120) 및 음성 구간 검출부(130)를 포함할 수 있다.
- [0028] 입술 영역 검출부(110)는 영상 신호로부터 사용자의 입술이 존재하는 영역을 검출할 수 있다. 입술 영역을 검출하는 방법과 관련하여, 도 2를 참조하여, 더 자세히 설명하기로 한다.
- [0029] 도 2는 일 실시 예에 있어서, 입술 영역을 검출하는 일 예를 도시한 예시도이다.
- [0030] 도 2의 210에 도시된 바와 같이, 입술 영역 검출부(110)는 영상에서 얼굴 영역을 검출하고, 얼굴 영역에서 눈을 검출하여, 눈과 입의 기하학적 위치 정보를 기반으로 입술 영역을 검출할 수 있다. 얼굴 영역의 검출은 Viola and Jones가 제안한 얼굴 검출 방법 또는 다양한 얼굴 검출 방법을 이용할 수 있다.

- [0031] 또 다른 방법으로, 입술 영역 검출부(110)는 도 2의 220에 도시된 바와 같이, 얼굴 영역을 검출하고, 검출된 얼굴 영역에서, YCbCr 색상 공간을 활용하여 입술 영역을 검출할 수 있다. 220을 참조하면, 입술 영역 검출부(110)는 YCbCr 색상 공간에서 Cb, Cr 값이 뚜렷하게 대비되는 영역을 검출하고, 검출된 영역의 픽셀 값들을 이진화할 수 있다. 입술 영역 검출부(110)는 이진화된 영상 영역에서 입술 형태를 포함하는 영역을 입술 영역으로 검출할 수 있다. YCbCr 색상 공간에서 입술의 Cr 및 Cb 값이 얼굴의 Cr 및 Cb 값과 비교할 때 뚜렷하게 대비될 수 있다. 따라서, 입술 영역 검출부(110)는 얼굴 영역에서 Cr 및 Cb 값을 이용하여 입술을 검출할 수 있다.
- [0032] 상술된 예에 한하지 않고, 입술 영역 검출부(110)는 다양한 방법으로 영상에서 입술 영역을 검출할 수 있다.
- [0033] 한편, 도 1의 입술 영역 움직임 분석부(120)는 입술 영역의 특징을 분석하여, 사용자가 발화 중인 것으로 판단되는 영상에 기초하여 음성 구간을 검출할 수 있다. 입과 치아의 형태적인 특징으로 인하여 닫힌 입술을 포함하는 영상보다 벌린 입술을 포함하는 영상에서 주변 픽셀 간 픽셀값의 차이가 더 클 수 있다. 따라서, 입술 영역 움직임 분석부(120)는 주변 픽셀 간 픽셀값의 차이를 구함으로써 입술의 움직임을 검출할 수 있다.
- [0034] 입술 영역 움직임 분석부(120)는 픽셀 값을 이용하여 영상을 분석하기 위해 입술 영역의 영상을 그레이 스케일 영상으로 변환하여 입술 영역의 특징을 분석할 수 있다.
- [0035] 입술 영역 움직임 분석부(120)는 이하 수학식 1을 이용하여 영상의 입술 영역에서 기준 픽셀과 기준 픽셀의 주변에 위치한 주변 픽셀들의 분산값을 구할 수 있다. 기준 픽셀은 입술 영역의 영상에 포함된 픽셀 중 적어도 하나일 수 있다.

수학식 1

$$VAR_{p,r} = \frac{1}{P} \sum_{p=0}^P (g_p - \mu)^2, \text{ where } \mu = \frac{1}{P} \sum_{p=0}^P g_p$$

- [0036]
- [0037] g_p 는 기준 픽셀의 위치에서 R 만큼 등거리에 떨어진 P개의 픽셀들의 픽셀 값 및 기준 픽셀의 픽셀값을 가리킬 수 있다. 예를 들면, R값은 1, P값은 8로 설정되는 경우, g_p 는 기준 픽셀로부터 1만큼 떨어진 8개의 픽셀 값들과 기준 픽셀의 픽셀 값을 가리킬 수 있다.
- [0038] $VAR_{p,r}$ 값은 평균값 μ 와 주변 픽셀들의 픽셀값 g_p 의 차이를 이용하므로, 영상의 조도 변화에 불변하는 특성이 있다. 따라서, 차량 내부 환경과 같이 조도가 변화하는 야외 환경에서 $VAR_{p,r}$ 값은 강인한 특징을 가질 수 있다. 또한 얼굴이 좌우로 움직이는 경우에도 픽셀값의 차이에 기초한 $VAR_{p,r}$ 값은 불변하므로, 얼굴 움직임에 따른 입술 영역 특징값의 오검출이 최소화 될 수 있다.
- [0039] 이하 도 3을 참조하여, $VAR_{p,r}$ 값을 구하는 방법에 관하여 더 자세히 설명하기로 한다.
- [0040] 도 3은 일 실시 예에 의한 $VAR_{p,r}$ 값을 구하는 방법의 일 예를 나타내는 예시 도면이다.
- [0041] 도 3을 참조하면, 입술 영역 움직임 분석부(120)는 입술 영역에 포함된 각 픽셀을 기준으로 3 by 3 크기의 블록 내에서 $VAR_{p,r}$ 값을 구할 수 있다. 기준 픽셀(301)은 블록 중심에 위치할 수 있다. 입술 영역 움직임 분석부(120)는 입술 영역의 각 픽셀을 기준으로 구한 $VAR_{p,r}$ 값에 대한 히스토그램을 구할 수 있다. 입술 영역의 영상이 히스토그램 정규화를 위해 60 픽셀 by 40 픽셀로 리사이즈된 후, $VAR_{p,r}$ 값에 대한 히스토그램이 획득될 수 있다. $VAR_{p,r}$ 값에 대한 히스토그램은 이하에서, LVH(Local Variance Histogram)로 나타내기로 한다.
- [0042] LVH에 관한 설명은 이하 도 4를 참조하여 더 자세히 설명하기로 한다.
- [0043] 도 4는 일 실시 예에 의한 LVH의 일 예를 나타낸 예시 도면이다.
- [0044] 도 4의 410 및 420을 참조하면, 닫힌 입을 포함하는 영상은 열린 입을 포함하는 영상에 비해, 주변 픽셀 간의 픽셀 값 차이가 적다. 따라서, 닫힌 입을 포함하는 영상의 LVH(410)에 비해 열린 입을 포함하는 영상의 LVH(420)는 상대적으로 큰 $VAR_{p,r}$ 값을 갖는 픽셀을 더 많이 포함할 수 있다. 따라서, 입술 영역 움직임 분석부(120)는 LVH에서 큰 $VAR_{p,r}$ 값을 갖는 픽셀 수에 기초하여, 각 영상에서의 입술 영역에 대한 특징값을 획득할 수

있다.

[0045] 입술 영역 움직임 분석부(120)는 수학식 2와 같이 각 영상의 입술 영역에 대한 LVH에서, 임계값인 T보다 $VAR_{p,R}$ 값이 큰 픽셀의 개수를 구할 수 있다.

[0046] 더하여, 발화 중에는 입술이 계속 움직이므로, 인접한 영상 간 입술 형태가 크게 변할 수 있는 점을 고려하여, 입술 영역 움직임 분석부(120)는 인접한 영상과의 픽셀 수 차이를 입술 영역에 대한 특징값으로 구할 수 있다.

수학식 2

$$g[n] = \sum_{i=T}^M H_n(i), \chi[n] = |g[n] - g[n-1]|$$

[0047]

[0048] $H_n(i)$ 는 n번째 영상의 입술 영역에서 $VAR_{p,R}$ 값이 i인 픽셀 수를 의미하고, i는 0부터 M의 값을 가질 수 있다. $g(n)$ 은 n번째 영상에서 $VAR_{p,R}$ 값이 임계값 T보다 큰 픽셀 수의 합을 의미한다. $x(n)$ 은 n번째 영상과 인접한 영상 간 g 값의 차이를 의미한다.

[0049] 이하 도 5를 참조하여, $x(n)$ 값에 관하여 더 자세히 설명하기로 한다.

[0050] 도 5는 일 실시 예에 의한 $x(n)$ 값을 도시한 그래프를 나타낸 도면이다.

[0051] 발화 중에는 입술이 계속 움직이므로, 인접한 영상 간 입술 형태가 크게 변할 수 있다. 따라서, 발화 중에는 인접한 영상 간 임계값보다 큰 $VAR_{p,R}$ 값을 갖는 픽셀 수의 차이가 커질 수 있으므로, $x(n)$ 은 비음성 구간보다 음성 구간에서 더 큰 값을 가질 수 있다.

[0052] 도 5를 참조하면, 음성 구간인 501, 502, 503 구간에서 $x(n)$ 값이 비음성 구간의 $x(n)$ 값에 비해 더 큰 값을 가진다. 따라서, 디바이스(100)는 $x(n)$ 값에 기초하여 음성 구간 및 비음성 구간을 구분할 수 있다.

[0053] 한편, 임계값 T는 음성 신호와 잡음신호에 대한 통계적인 모델에 기초하여 결정될 수 있다. 현재의 입력 영상이 음성 구간이라는 가설(hypothesis)을 H_1 , 비음성 구간이라는 가설을 H_0 라고 하면, 수학식 3과 같이 두 가설을 나타낼 수 있다.

수학식 3

$$H_0(\text{no lip activity}): \chi[n] = \omega[n]$$

[0054]

$$H_1(\text{lip activity}): \chi[n] = s[n] + \omega[n]$$

[0055]

[0056] $s(n)$ 과 $\omega(n)$ 은 각각 n번째 영상의 에너지에 대한 신호와 잡음 신호를 의미한다. 입술 영역 움직임 분석부(120)는 잡음 신호 $\omega(n)$ 를 모델링하기 위하여, 처음 일정 구간의 영상 구간은 항상 비음성 구간이라 가정할 수 있다.

[0057] 일 실시 예에 의한 입술 영역 움직임 분석부(120)는 임계값 T를 구하기 위하여, 네이만-피어슨(neyman-pearson) 이론을 이용하여 $P_{FA} = \alpha$, $\alpha = 0.1$ 로 두고 임계값 T를 구할 수 있다. α 값은 오차의 허용한계를 나타내는 값으로, 0.1 이외에 다른 값을 가질 수 있다.

수학식 4

$$\Lambda(X) = \frac{p(X|H_1)}{p(X|H_0)} > T$$

[0058]

[0059] 수학식 4에서 $X = [x(1), x(2), \dots, x(N)]$ 인 벡터이고, $\Lambda(X)$ 는 음성 및 비음성 구간에 대한 우도비

(likelihood ratio)를 의미한다. $p(X|H_1)$ 은 입력영상이 음성 구간이라는 가설의 우도(likelihood)이며, $p(X|H_0)$ 은 입력영상이 비음성 구간이라는 가설의 우도이다. 우도비가 임계값 T보다 큰 경우 입력 영상이 음성 구간이라는 가설이 채택될 수 있다. 이하 수학식 5를 통해서 사용자가 지정한 오차 허용한계에 따른 임계값 T가 결정될 수 있다.

수학식 5

$$P_{FA} = \int_T^{\infty} p(X|H_0) dX = \alpha$$

[0060]

[0061]

입술 영역 움직임 분석부(120)는 수학식 5를 통해서, $x(n)$ 값을 구하기 위한 임계값 T를 결정할 수 있다. 비음성 구간에 속하는 입력영상이 음성구간으로 잘못 판단되는 경우의 확률인 P_{FA} 를 오차의 허용한계 α 값으로 만들어주는 T를 추정함으로써 임계값 T를 결정할 수 있다. 입술 영역 움직임 분석부(120)는 각 영상에 대하여 $x(n)$ 을 획득하고, 최종적으로 음성 구간을 검출하기 위하여 에지 검출 필터(edge detection filter)를 이용할 수 있다. 입술 영역 움직임 분석부(120)는 음성 시작 구간에서는 특징값인 $x(n)$ 이 커지고, 음성이 끝나는 구간에서는 특징값 $x(n)$ 이 감소하는 특징을 이용하기 위한 에지 검출 필터를 특징값에 적용할 수 있다. 그리고, 음성 구간 검출부(130)는 에지 검출 필터를 적용한 특징값 및 상태 전이 모델(state transition model)을 이용하여 최종적으로 음성 구간을 검출할 수 있다.

[0062]

에지 검출 필터에 관하여, 도 6을 참조하여 더 자세히 설명하기로 한다.

[0063]

도 6은 일 실시 예에 의한 에지 검출 필터의 일 예를 그래프로 나타낸 예시 도면이다.

[0064]

에지 검출 필터 $h(i)$ 는 수학식 6과 같이 표현될 수 있다.

수학식 6

$$f(\chi) = e^{A_1} [K_1 \sin(A\chi) + K_2 \cos(A\chi)] + e^{-A_1} [K_3 \sin(A\chi) + K_4 \cos(A\chi)] + K_5 + K_6 e^{-s\chi}$$

[0065]

$$h(i) = \begin{cases} -f(i), & -W \leq i \leq 0 \\ f(i), & 1 \leq i \leq W \end{cases}$$

[0066]

[0067]

W는 필터 길이와 관계되는 변수이며, i는 -W부터 W까지 정수값을 가질 수 있다. A(A=0.41)와 K(K₁=1.538, K₂=1.468, K₃=-0.078, K₄=-0.036, K₅=-0.872, K₆=-0.56)는 필터 파라미터이다. 입술 영역 움직임 분석부(120)는 특징값 $x(n)$ 에 에지 검출 필터 $h(i)$ 를 적용하여, 출력 F(n)을 이하 수학식 7과 같이 획득할 수 있다. F(n)은 영상의 입술 움직임에 대한 특징값인 $x(n)$ 에 에지 검출 필터를 적용한 결과를 의미한다.

수학식 7

$$F(n) = \sum_{i=-W}^W h(i) x(n+i)$$

[0068]

[0069]

에지 검출 필터 $h(i)$ (W=7)는 도 6에 도시된 그래프와 같이 원점에 대칭된다. 비음성 구간에서는 $x(n)$ 값이 그 크기와는 관계없이 일정하므로 $h(i)x(n+i)$ 값이 -W부터 W까지 동일한 크기의 양수 값 및 음수 값을 가짐에 따라 에지 검출 필터가 적용된 특징값이 0에 가까울 수 있다. 반면에, 음성 구간에서는 $x(n)$ 값이 커지게 되어 양수 값을 가진 성분이 더 커짐에 따라 에지 검출 필터가 적용된 특징값이 커지게 된다. 반면에, 특징값이 작아지면 음수 값을 가진 성분이 더 커짐에 따라 에지 검출 필터가 적용된 특징값이 작아지게 된다. 따라서, 입술 영역 움직임 분석부(120)는 에지 검출 필터를 이용하여 잡음의 크기에 따라 문턱치를 조절할 필요 없이 서서히 변하는 특징값에도 강인하게 음성 구간을 검출할 수 있다.

- [0070] 음성 구간 검출부(130)는 $F(n)$ 값과 상태 천이 모델(state transition model)을 이용하여 최종적으로 음성 구간을 검출할 수 있다.
- [0071] 상태 천이 모델을 이용하여 음성 구간을 검출하는 방법에 대하여, 이하 도 7 내지 도 9를 참조하여 더 자세히 설명하기로 한다.
- [0072] 도 7은 일 실시 예에 의한 상태 천이 모델의 일 예를 나타낸 예시 도면이다.
- [0073] 도 7을 참조하면, 음성 구간 검출부(130)는 도 7에 도시된 상태 천이 모델(State transition model)을 이용하여 음성 구간을 검출할 수 있다.
- [0074] 비음성 구간(701)은 영상과 대응되는 음향 신호에 음성이 존재하지 않는 것으로 판단되는 구간을 의미한다. 음성 구간(702)은 영상과 대응되는 음향 신호에 음성 구간 검출부(130)에서 검출하고자 하는 음성이 존재하는 것으로 판단되는 구간을 의미한다. 묵음 구간(703)은 음성 구간(702)이나 후술될 카운트 값에 따라 비음성 구간(701)으로 상태가 변환될 수 있는 구간을 의미한다. T_L , 즉, 낮은 임계값(lower threshold), T_U , 즉, 높은 임계값(upper threshold) 및 gap은 음성 구간(702)의 종료 여부를 결정하기 위한 기준 값으로, 실험적으로 결정될 수 있는 상수이다. T_U 는 항상 T_L 보다 큰 값을 가질 수 있다.
- [0075] 음성 구간 검출부(130)는 상태 천이 모델을 이용하여, $F(n)$ 이 T_U 보다 작으면 음성이 없는 비음성 구간(701)으로 판단할 수 있다. $F(n)$ 이 T_U 보다 커지면 음성 구간(702)이 시작된 것으로 보고 음성 구간 검출부(130)는 $F(n)$ 이 T_U 보다 커진 지점을 음성 구간(702)이 시작된 지점으로 결정할 수 있다. 음성 구간(702)이 시작된 이후, $F(n)$ 이 T_L 보다 작아지면, 음성 구간 검출부(130)는 묵음 구간(703)으로 상태가 변환된 것으로 보고, count 값을 0으로 설정할 수 있다.
- [0076] 그리고, 음성 구간 검출부(130)는 $F(n)$ 이 T_U 보다 작으나, T_L 보다 큰 상태에 있는 동안, 묵음 구간(703)에 있는 것으로 보고, 시간이 경과함에 따라 count 값을 증가시킬 수 있다. 음성 구간 검출부(130)는 count 값이 미리 설정된 gap 값보다 커지게 되면, 음성 구간(702)은 종료된 것으로 보고, 현재 영상을 비음성 구간(701)에 있는 것으로 판단할 수 있다. 한편, 음성 구간 검출부(130)는 $F(n)$ 값이 T_U 보다 커지면, 음성 구간(702)으로 상태를 변환시킬 수 있다.
- [0077] 도 8은 일 실시 예에 의한 $x(n)$, $F(n)$ 및 음성 구간으로 검출된 구간의 일 예를 나타낸 그래프이다.
- [0078] 도 8의 810을 참조하면, LVH의 특징값으로 추출될 수 있는 $x(n)$ 값이 도시되어 있다. 830은 음성 구간으로 검출된 구간을 나타낸 그래프이다. 음성 구간으로 검출된 구간인 830 그래프와 810의 그래프를 비교해볼 때, 상술한 바와 같이 음성 구간에 해당하는 구간에서, $x(n)$ 값의 크기가 크게 변할 수 있다.
- [0079] 820을 참조하면, $x(n)$ 에 대해 에지 검출 필터를 적용한 결과인 $F(n)$ 을 나타낸 그래프가 도시되어 있다. 상술된 T_L 값보다 $F(n)$ 값이 크거나, T_U 값보다 $F(n)$ 값이 작은 값을 갖는 구간과 830에 도시된 음성 구간과 대응될 수 있다. 또한, 830에 도시된 음성 구간과 대응되는 구간에서, $F(n)$ 값이 T_L 값과 T_U 값 사이에 존재하는 묵음 구간이 존재한다. 묵음 구간은 비음성 구간에 비하여 상대적으로 작은 시간동안에 존재하므로, 음성 구간 검출부(130)는 count 값이 gap 값보다 작은 값을 가지는 구간은, 음성 구간으로 판단할 수 있다.
- [0080] 도 9는 일 실시 예에 의한 묵음 구간의 일 예를 나타낸 예시 도면이다.
- [0081] 도 9의 910을 참조하면, '스마트폰'이란 음성이 발화될 때, '트'와 '폰' 사이의 일정 시간 동안 발생되지 않는 구간인 gap 구간(911)이 존재할 수 있다. gap 구간(911) 동안에는 발생이 되지 않음에 따라 입술의 움직임 변화가 둔화될 수 있다. 따라서, 영상의 입술 움직임 특징값으로 검출될 수 있는 $x(n)$ 값도 gap 구간(921)에서 0에 가까운 값을 가질 수 있다.
- [0082] 음성 구간 검출부(130)는 상술된 상태 천이 모델을 이용하여 음성 구간을 결정할 때 이용할 수 있는 gap 구간의 길이를 실험적으로 결정할 수 있다. 예를 들면, 음성 구간 검출부(130)는 사용자별 발화습관 또는 발화습관에 영향을 줄 수 있는 사용자의 특징 정보에 따라 gap 구간의 길이를 결정할 수 있다.
- [0083] 이하 도 10을 참조하여, 디바이스(100)에 의하여 음성 구간을 검출하는 방법에 대해 자세히 설명하기로 한다.
- [0084] 도 10은 일 실시 예에 의한 음성 구간을 검출하는 방법을 나타낸 순서도이다.

- [0085] 도 10을 참조하면, 단계 S1001에서, 디바이스(100)는 영상에서 입술이 존재하는 영역을 검출할 수 있다. 예를 들면, 디바이스(100)는 얼굴 영역을 검출하고, 얼굴 영역에서 눈을 검출하여, 눈과 입의 기하학적 위치 정보를 기반으로 입술 영역을 검출할 수 있다. 또 다른 방법으로, 디바이스(100)는 얼굴 영역을 검출하고, 검출된 얼굴 영역에서, YCbCr 색상 공간에서의 픽셀 값 차이에 기초하여 입술 영역을 검출할 수 있다.
- [0086] 단계 S1003에서, 디바이스(100)는 검출된 영역에 포함된 각 픽셀에 대한 주변 픽셀들 간의 픽셀값의 차이에 기초하여, 입술 움직임에 대한 특징값을 획득할 수 있다.
- [0087] 예를 들면, 디바이스(100)는 검출된 입술 영역에 포함된 각 픽셀을 기준으로 하여, 주변 픽셀들 간의 픽셀값의 차이를 구하고, 차이값에 기초하여 입술 영역에 대한 특징 값을 획득할 수 있다. 입술 영역의 특징값은 기준 픽셀과 주변 픽셀들의 분산값 $VAR_{p,R}$ 에 기초하여 획득될 수 있다. 입술 움직임에 대한 특징값은 임계값으로 설정된 T 값보다 큰 분산 값을 가지는 픽셀들의 개수를 나타내는 LVH에 기초하여 결정될 수 있다. LVH에 기초하여 결정된 픽셀 수를 LVH 값으로 나타내기로 한다.
- [0088] 더하여, 디바이스(100)는 발화 중일 때, 입술이 짧은 시간동안 계속 닫혔다 열리는 동작을 지속하는 점을 고려하여, 인접한 영상간 LVH 값의 차이값을 입술 움직임의 특징값으로 결정할 수 있다.
- [0089] 또한, 디바이스(100)는 입술 움직임의 특징값에 대하여, 에지 검출 필터를 적용함으로써, 잡음으로 인한 에러 발생을 최소화할 수 있다.
- [0090] 단계 S1005에서, 디바이스(100)는 단계 S1003에서 획득한 특징값에 기초하여 음성 구간을 검출할 수 있다. 예를 들면, 디바이스(100)는 상술한 상태 천이 모델을 이용하여 음성 구간을 검출할 수 있다.
- [0091] 디바이스(100)는 T_L 값보다 특징값이 커지는 시점을 음성 구간으로 검출할 수 있고, 음성 구간에서, 특징값이 T_U 값과 T_L 값 사이에 존재하는 동안 count 값을 증가시킬 수 있다. 그리고, 디바이스(100)는 count 값이 미리 설정된 gap 값보다 커지면 대응되는 영상 구간을 비음성 구간으로 결정할 수 있다.
- [0092] 일 실시 예에서, 음성 구간으로 검출될 수 있는 영상 구간은 영상의 입술 영역에 대한 특징값을 기초로 검출된 것으로, 엄밀하게는 입술 영역의 특징에 기초하여 발화 중일 확률이 높은 구간으로 추정된 구간이다. 디바이스(100)는 실제 음향 신호를 참조하여, 영상의 입술 영역의 특징값에 기초하여 검출된 음성 구간을 수정 변환할 수 있다.
- [0093] 이하 도 11 내지 도 12를 참조하여, 입술 움직임에 기초하여 음성 구간을 검출하는 또 다른 방법에 대하여 더 자세히 설명하기로 한다.
- [0094] 도 11은 일 실시 예에 의한 입술 움직임에 기초하여 음성 구간을 검출하는 방법을 나타낸 순서도이다.
- [0095] 도 11을 참조하면, 단계 S1101에서, 디바이스(100)는 영상에서 입술이 존재하는 영역을 검출할 수 있다. 예를 들면, 디바이스(100)는 얼굴 영역을 검출하고, 얼굴 영역에서 눈을 검출하여, 눈과 입의 기하학적 위치 정보를 기반으로 입술 영역을 검출할 수 있다. 또 다른 방법으로, 디바이스(100)는 얼굴 영역을 검출하고, 검출된 얼굴 영역에서, YCbCr 색상 공간에서의 픽셀 값에 기초하여 입술 영역을 검출할 수 있다.
- [0096] 단계 S1103에서, 디바이스(100)는 단계 S1101에서 검출된 입술 영역에 포함된 각 픽셀에 대하여 움직임 정보를 획득할 수 있다. 각 픽셀의 움직임 정보는 이하 수식식 8과 같이 각 픽셀에 대한 움직임 벡터값을 포함할 수 있다. $v_x(i, j)$ 와 $v_y(i, j)$ 는 $m \times n$ 사이즈의 입술 영역의 영상의 픽셀 (i, j) 위치에서 x, y 방향에 대한 속도 값을 나타낸다.

수학식 8

- [0097]
$$V=[v_x(1,1),\dots,v_x(m,n),v_y(1,1),\dots,v_y(m,n)]$$

- [0098] 예를 들면, 디바이스(100)는 인접 영상의 픽셀값을 이용하여 현재 영상의 각 픽셀에 대한 움직임 정보를 획득할 수 있다. 디바이스(100)는 연속된 영상에서 각 픽셀의 움직임을 속도 벡터로 나타내기 위하여 옵티컬 플로우(optical flow) 방법을 이용할 수 있다. 옵티컬 플로우 방법을 이용하여, 디바이스(100)는 연속된 두 영상에서 가장 비슷한 블록을 찾음으로써, 각 픽셀의 움직임을 속도 벡터로 획득할 수 있다.

[0099] 디바이스(100)는 옵티컬 플로우를 이용하여 각 픽셀의 움직임 정보를 획득하는 경우, 시간상으로 인접한 두 영상에서의 픽셀값을 기준으로 움직임 정보를 획득할 수 있다. 따라서, 디바이스(100)는 조도 변화가 다양한 환경에서도 강인하게 움직임 정보를 획득할 수 있다.

[0100] 단계 S1105에서, 디바이스(100)는 단계 S1101에서 검출된 입술 영역에 대하여 분할된 영역이 서로 대칭되도록 분할할 수 있다.

[0101] 입술은 상하 또는 좌우 방향으로 대칭되는 특성이 있으며, 입술의 움직임 역시 상하 또는 좌우 방향으로 대칭된다는 특성이 있다. 입술이 고정되어 있고, 얼굴이 움직이는 상황에서 입술이 움직이지 않음에도 입술 영역에 대해 움직임 정보가 존재하므로, 입술 움직임에 대한 특징값 추출 시 에러가 발생할 수 있다. 따라서, 디바이스(100)는 입술 움직임 정보의 대칭성을 이용하여 입술 움직임을 특징값을 구할 수 있다.

[0102] 단계 S1107에서, 디바이스(100)는 단계 S1105에서 분할된 각 영역에서의 움직임 정보가 차이값을 이용하여 입술 움직임을 특징값을 획득할 수 있다. 예를 들면, 디바이스(100)는 수식식 9와 같이 입술 움직임을 특징값 E를 획득할 수 있다.

수식식 9

$$E = \left| \sum_{i,j \in A} v_x(i,j) - \sum_{i,j \in B} v_x(i,j) \right|^2 + \left| \sum_{i,j \in A} v_y(i,j) - \sum_{i,j \in B} v_y(i,j) \right|^2$$

[0104] 입술 움직임을 특징값 E는 분할된 영역에 각각 포함된 움직임 벡터 값의 차이에 기초하여 획득될 수 있다. 따라서, 특징값 E는 얼굴 움직임과 같이 각 영역에 포함된 동일 방향의 움직임 벡터 값은 상쇄되고, 입술 움직임과 같이 반대 방향의 움직임 벡터 값만 존재할 수 있다.

[0105] 단계 S1109에서, 디바이스(100)는 단계 S1107에서 획득된 특징값에 기초하여 음성 구간을 검출할 수 있다.

[0106] 예를 들면, 디바이스(100)는 상술된 x(n) 값과 같이 인접 영상과의 특징값 차이를 입술 움직임을 특징값으로 재결정할 수 있다. 또한, 디바이스(100)는 입술 움직임을 특징값에 대하여, 에지 검출 필터를 적용함으로써, 잡음으로 인한 에러 발생을 최소화할 수 있다. 그리고, 디바이스(100)는 상술한 상태 천이 모델을 이용하여 음성 구간을 검출할 수 있다.

[0107] 도 12는 일 실시 예에 의한 입술 영역을 분할하는 일 예를 나타낸 예시 도면이다.

[0108] 도 12의 1210을 참조하면, 디바이스(100)는 좌우 방향으로 입술 영역을 분할하고, 각 픽셀의 x방향의 벡터 값 간 차이에 기초하여 특징값 E를 구할 수 있다. 또한, 1220을 참조하면, 디바이스(100)는 상하 방향으로 입술 영역을 분할하고, 각 픽셀의 y방향의 벡터 값 간 차이에 기초하여 특징값 E를 구할 수 있다.

[0109] 도 13은 일 실시 예에 의한 입술 움직임에 기초하여 음성 구간을 검출하는 디바이스의 내부 구조를 나타낸 블록도이다.

[0110] 일 실시 예에 의한 디바이스(1300)는 음성 인식을 수행할 수 있는 단말 장치일 수 있으며, 예를 들면, 스마트 TV(television), UHD(ultra high definition) TV, 모니터, PC(Personal Computer), 노트북 컴퓨터, 휴대폰(mobile phone), 태블릿 PC, 내비게이션(navigation) 단말기, 스마트폰(smart phone), PDA(Personal Digital Assistants), PMP(Portable Multimedia Player) 및 디지털방송 수신기를 포함할 수 있다.

[0111] 도 13을 참조하면, 디바이스(1300)는 수신부(1310) 및 제어부(1320)를 포함할 수 있다.

[0112] 수신부(1310)는 음성 구간을 검출하기 위한 사용자의 입술 영역을 포함하는 영상을 획득할 수 있다. 수신부(1310)는 디바이스(1300)에 구비된 촬영 장치 또는 외부 장치로부터 입술 영역을 포함하는 영상을 획득할 수 있다. 더하여, 수신부(1310)는 디바이스(1300)에서 검출된 음성 구간에 기초하여 음성 인식을 수행하기 위한 사용자의 음성을 포함하는 음향 신호를 더 획득할 수 있다. 디바이스(1300)는 입술 영역을 포함하는 영상에 기초하여 음성 구간이 존재할 것으로 판단되는 구간을 추정하고, 추정된 음성 구간에 기초하여, 음향 신호에 포함된 사용자의 음성에 대하여 음성 인식을 수행할 수 있다.

[0113] 제어부(1320)는 수신부(1310)에 의해 수신된 입술 영역이 포함된 다수의 영상에 기초하여 음성 구간을 검출할 수 있다. 예를 들면, 제어부(1320)는 영상으로부터 입술 영역을 검출하고, 입술 영역 내 픽셀들 간 픽셀값의 차이에 기초하여 각 영상의 입술 움직임에 대한 특징값을 획득할 수 있다. 제어부(1320)는 주변 픽셀들 간의 픽셀

값의 차이에 기초하여 획득한 각 픽셀을 기준으로 한 분산값을 구하고, 임계값으로 설정된 T 값보다 큰 분산 값을 가지는 픽셀들의 개수인 LVH 값을 구할 수 있다.

[0114] 더하여, 제어부(1320)는 발화 중일 때, 입술이 짧은 시간동안 계속 닫혔다 열리는 동작을 지속하는 점을 고려하여, 인접한 영상간 LVH 값의 차이값을 입술 움직임의 특징값으로 결정할 수 있다.

[0115] 또는, 제어부(1320)는 입술 영역에 포함된 각 픽셀에 대한 움직임 정보에 기초하여 입술 움직임에 대한 특징값을 획득할 수 있다. 제어부(1320)는 입술 움직임에 대한 특징값에 기초하여, 상태 천이 모델을 이용하여 영상으로부터 음성 구간을 검출할 수 있다. 각 픽셀에 대한 움직임 정보는 옵티컬 플로우 방법에 따라 획득될 수 있다. 제어부(1320)는 입술 영역이 서로 대칭되도록 분할된 각 영역에서의 움직임 정보의 차이값에 기초하여 입술 움직임에 대한 특징값을 획득할 수 있다.

[0116] 더하여, 제어부(1320)는 검출된 음성 구간과 대응되는 음향 신호에 대하여 음성 인식을 수행할 수 있다. 제어부(1320)는 사용자의 음성이 포함되어 있는 것으로 판단되는 음향 신호에 대하여 음성 인식을 수행함으로써, 잡음으로 인한 오인식을 최소화할 수 있다.

[0117] 일 실시 예에 의하면, 실내 조명 변화 및 얼굴 움직임과 같은 환경 변화에 강인한 입술 움직임의 특징값에 기초하여 영상에 기반한 음성 구간을 검출할 수 있다.

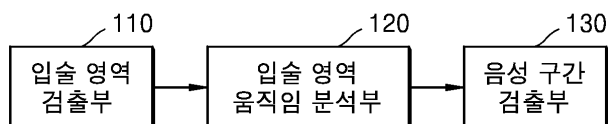
[0118] 일부 실시 예에 의한 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0119] 비록 상기 설명이 다양한 실시예들에 적용되는 본 발명의 신규한 특징들에 초점을 맞추어 설명되었지만, 본 기술 분야에 숙달된 기술을 가진 사람은 본 발명의 범위를 벗어나지 않으면서도 상기 설명된 장치 및 방법의 형태 및 세부 사항에서 다양한 삭제, 대체, 및 변경이 가능함을 이해할 것이다. 따라서, 본 발명의 범위는 상기 설명에서보다는 첨부된 특허청구범위에 의해 정의된다. 특허청구범위의 균등 범위 안의 모든 변형은 본 발명의 범위에 포함된다.

도면

도면1

100

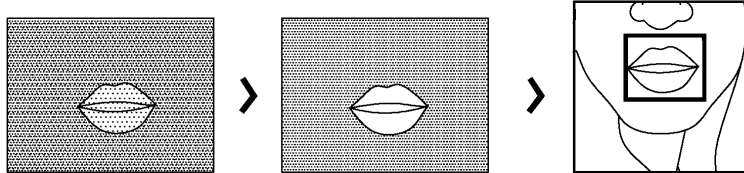


도면2

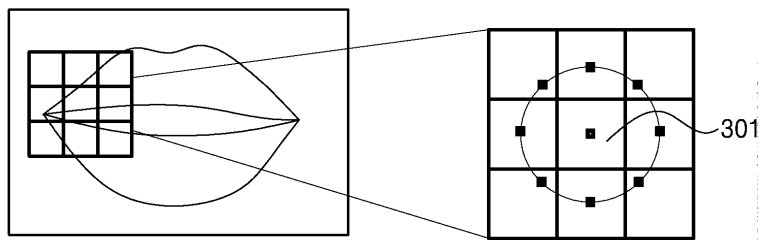
210



220

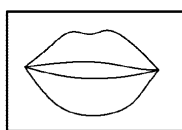


도면3

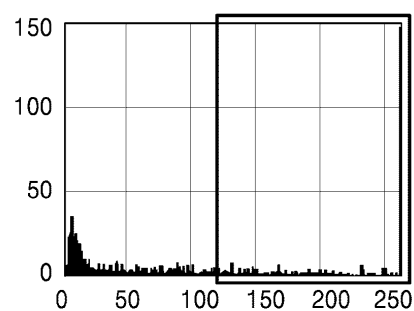
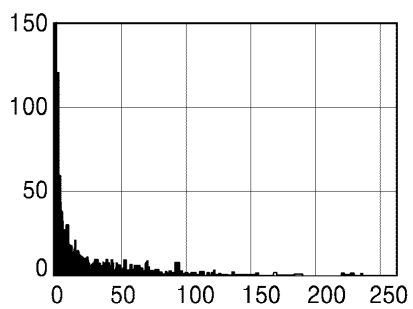
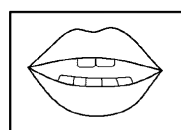


도면4

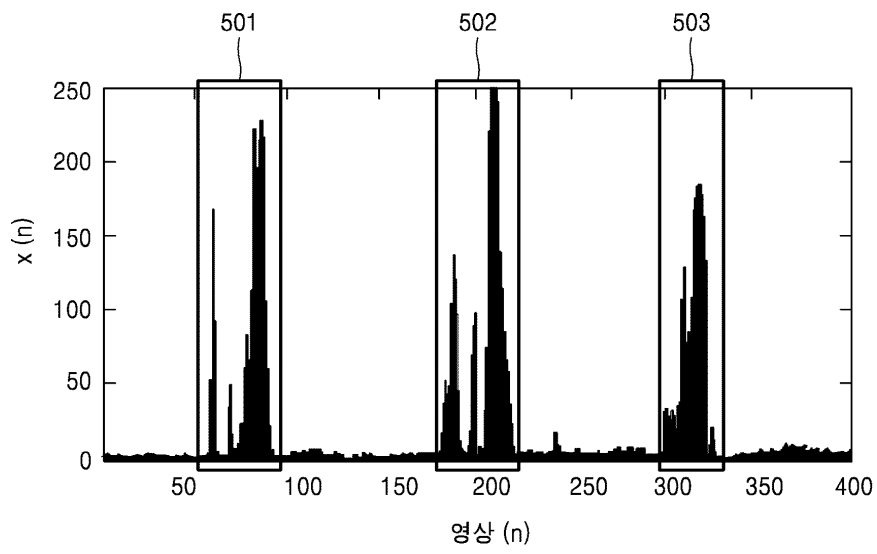
410



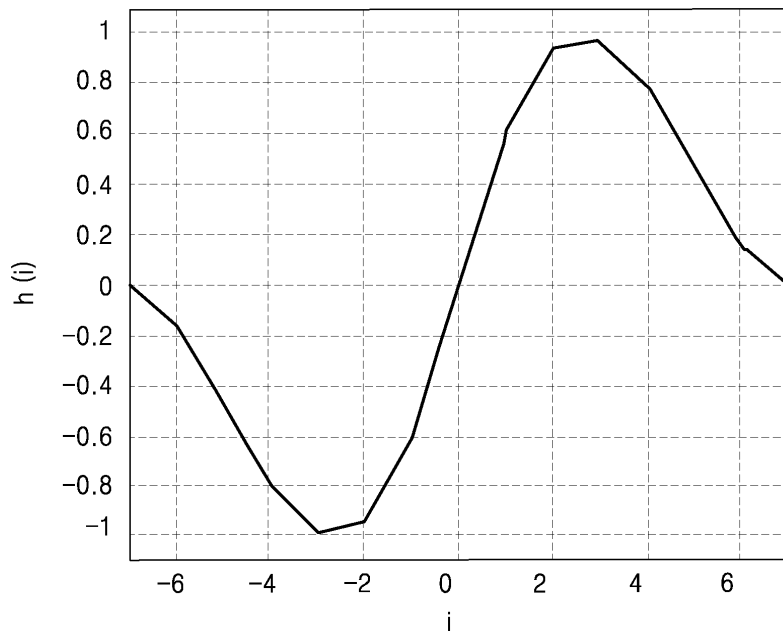
420



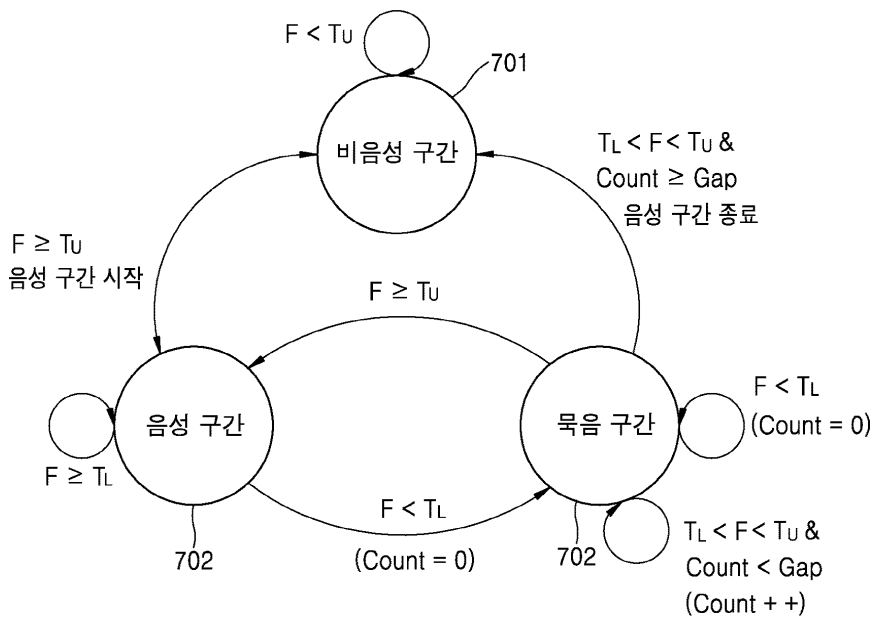
도면5



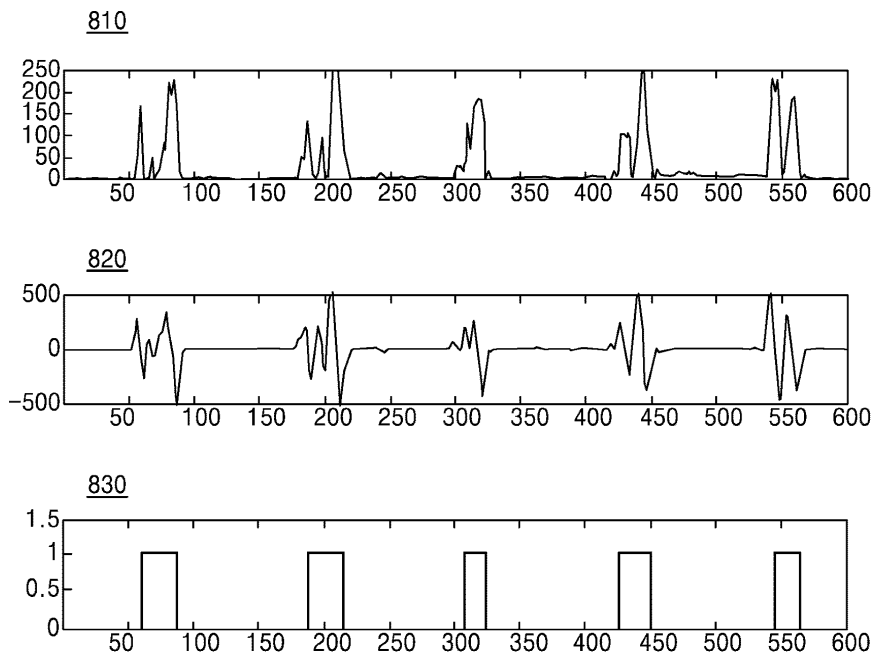
도면6



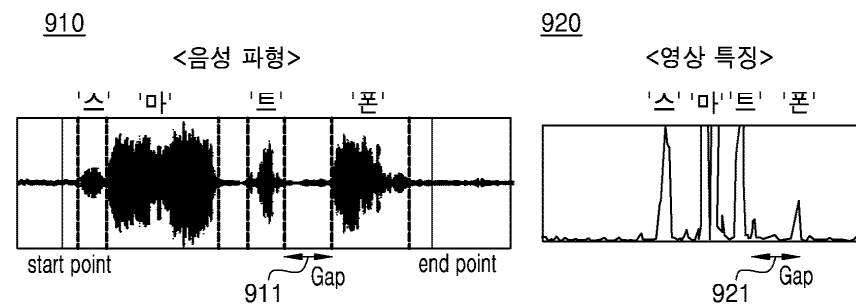
도면7



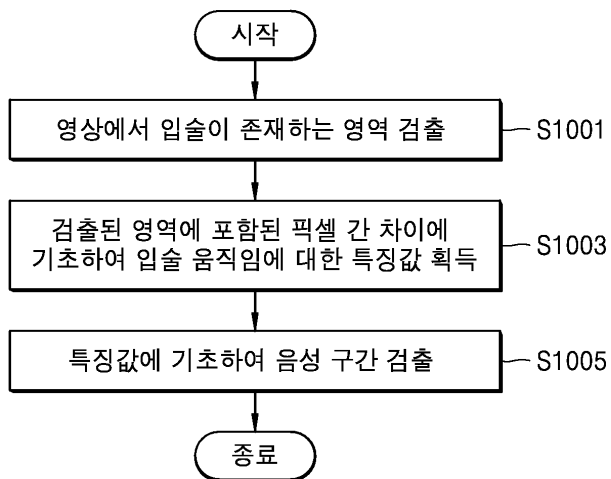
도면8



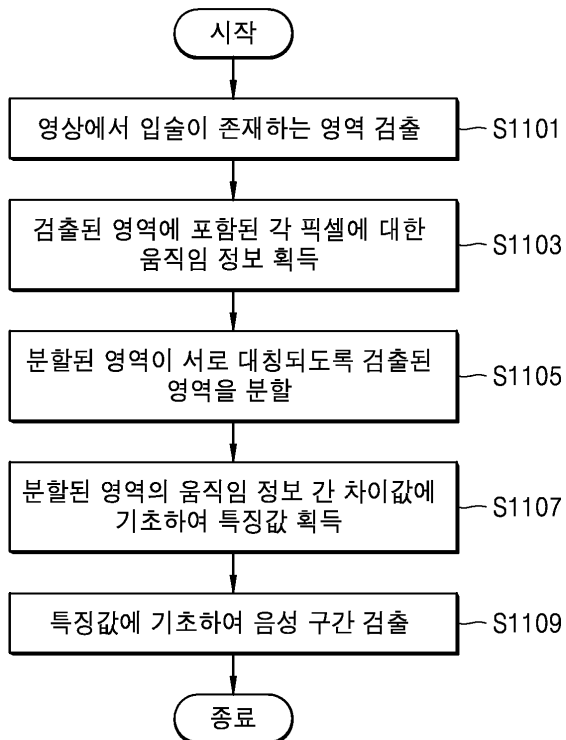
도면9



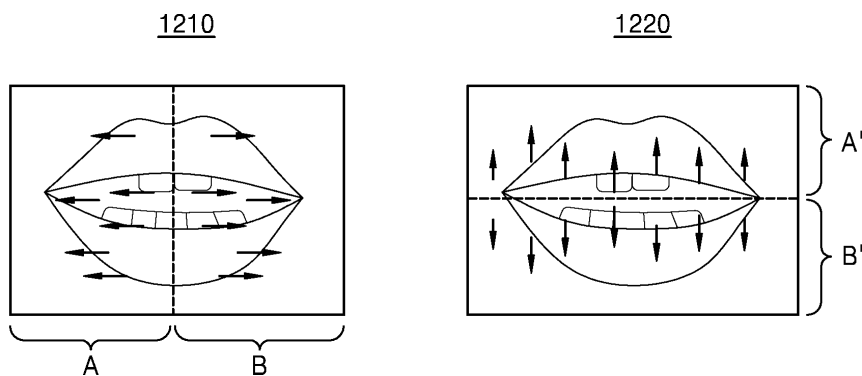
도면10



도면11



도면12



도면13

1300

