(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0142131
(43) 공개일자 2020년12월22일

(51) 국제특허분류(Int. Cl.)
*G10L 15/25* (2013.01) *G06N 20/00* (2019.01)
*G06T 3/40* (2006.01) *G06T 7/00* (2017.01)
(52) CPC특허분류
*G10L 15/25* (2013.01)
*G06N 20/00* (2019.01)
(21) 출원번호 10-2019-0068439
(22) 출원일자 2019년06월11일
심사청구일자 2019년06월11일

(71) 출원인
**서강대학교산학협력단**
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
**박형민**
서울특별시 강남구 삼성로 212, 3동 1002호
**박래홍**
서울특별시 영등포구 문래동3가 당산로 4길 12
*(뒷면에 계속)*
(74) 대리인
**이준영, 성원찬**

전체 청구항 수 : 총 17 항

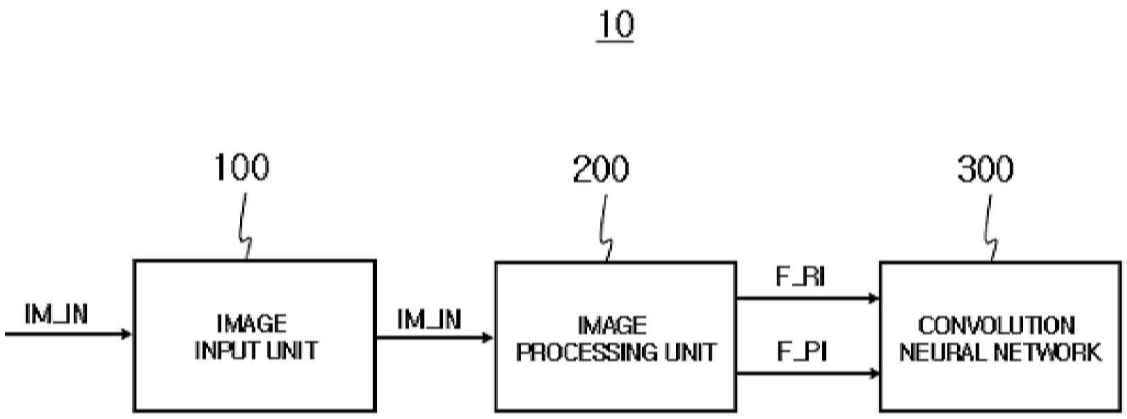
(54) 발명의 명칭 **음성인식 장치 및 음성인식 장치의 동작방법**

## (57) 요 약

본 발명의 실시예에 따른 음성인식 학습장치는 영상 입력부, 영상 처리부 및 컨볼루션 뉴럴 네트워크를 포함할 수 있다. 영상 입력부는 입력영상을 수신할 수 있다. 영상 처리부는 미리 정해진 제1 시간간격 동안의 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N

(뒷면에 계속)

*대 표 도* - 도1



은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크는 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 입술의 모양에 상응하는 음성정보를 학습할 수 있다.

본 발명에 따른 음성인식 장치에서는 화자의 입술영상 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

(52) CPC특허분류
*G06T 3/40* (2013.01)
*G06T 7/00* (2013.01)
*G06T 2207/20081* (2013.01)
*G06T 2207/20084* (2013.01)

(72) 발명자
**장동원**
서울특별시 강동구 성안로 162, 1102호
**제창수**
서울특별시 서대문구 통일로 319, 103동 1204호
**김홍인**
경상북도 문경시 돈달산길 1 904호 (모전동,성덕하이빌)

이 발명을 지원한 국가연구개발사업

| | |
|---|---|
| 과제고유번호 | 2017R1A2B4009964 |
| 부처명 | 과학기술정보통신부 |
| 과제관리(전문)기관명 | 재단법인 한국연구재단 |
| 연구사업명 | 이공분야기초연구사업 |
| 연구과제명 | 시청각 정보에 대한 멀티모달 딥러닝 기반의 강인한 연속음성인식 기술 개발 |
| 기 여 율 | 1/1 |
| 과제수행기관명 | 서강대학교 산학협력단 |
| 연구기간 | 2018.03.01 ~ 2019.02.28 |

# 명 세 서

## 청구범위

### 청구항 1

입력영상을 수신하는 영상 입력부;

미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나누는 영상 처리부; 및

상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성정보를 학습하는 컨볼루션 뉴럴 네트워크를 포함하는 음성인식 학습장치.

### 청구항 2

제1항에 있어서,

상기 영상 처리부는,

상기 제1 시간간격 동안의 상기 입술영상 및 상기 패치영상의 길이를 조절하여 N개의 프레임들로 나누는 라벨 변형기를 포함하는 것을 특징으로 하는 음성인식 학습장치.

### 청구항 3

제2항에 있어서,

상기 라벨 변형기는,

상기 제1 시간간격 동안의 상기 입술영상 및 상기 패치영상의 길이를 조절하여 길이조절 입술영상 및 길이조절 패치영상을 제공하는 영상길이 조절기; 및

상기 길이조절 입술영상 및 상기 길이조절 패치영상을 각각 N개의 프레임들로 나누어 상기 프레임 입술영상 및 상기 프레임 패치영상을 제공하는 프레임 영상 제공기를 포함하는 것을 특징으로 하는 음성인식 학습장치.

### 청구항 4

제3항에 있어서,

상기 입술영상의 길이를 조절하여 상기 입술영상의 길이가 증가하는 경우,

상기 패치영상의 길이를 조절하여 상기 패치영상의 길이가 증가되는 것을 특징으로 하는 음성인식 학습장치.

### 청구항 5

제3항에 있어서,

상기 입술영상의 길이를 조절하여 상기 입술영상의 길이가 감소하는 경우,

상기 패치영상의 길이를 조절하여 상기 패치영상의 길이가 감소되는 것을 특징으로 하는 음성인식 학습장치.

### 청구항 6

제3항에 있어서,

상기 컨볼루션 뉴럴 네트워크는 제1 컨볼루션 뉴럴 네트워크 및 제2 컨볼루션 뉴럴 네트워크를 포함하는 것을 특징으로 하는 음성인식 학습장치.

### 청구항 7

제6항에 있어서,

상기 제1 컨볼루션 뉴럴 네트워크는 상기 프레임 입술영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 학습하는 것을 특징으로 하는 음성인식 학습장치.

**청구항 8**

제6항에 있어서,

상기 제2 컨볼루션 뉴럴 네트워크는 상기 프레임 패치영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 학습하는 것을 특징으로 하는 음성인식 학습장치.

**청구항 9**

입력영상을 수신하는 영상 입력부;

미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나누는 영상 처리부; 및

상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성인식 결과를 결정하는 컨볼루션 뉴럴 네트워크를 포함하는 음성인식 장치.

**청구항 10**

제9항에 있어서,

상기 컨볼루션 뉴럴 네트워크는 제1 컨볼루션 뉴럴 네트워크 및 제2 컨볼루션 뉴럴 네트워크를 포함하는 것을 특징으로 하는 음성인식 장치.

**청구항 11**

제10항에 있어서,

상기 제1 컨볼루션 뉴럴 네트워크는 상기 프레임 입술영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 결정하는 것을 특징으로 하는 음성인식 장치.

**청구항 12**

제11항에 있어서,

상기 제2 컨볼루션 뉴럴 네트워크는 상기 프레임 패치영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 결정하는 것을 특징으로 하는 음성인식 장치.

**청구항 13**

제12항에 있어서,

상기 제1 컨볼루션 뉴럴 네트워크로부터 결정되는 음성정보에 상응하는 제1 확률 및 상기 제2 컨볼루션 뉴럴 네트워크로부터 결정되는 음성정보에 상응하는 제2 확률의 합에 기초하여 음성인식 결과가 결정되는 것을 특징으로 하는 음성인식 장치.

**청구항 14**

영상 입력부가 입력영상을 수신하는 단계;

영상 처리부가 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나누는 단계;

컨볼루션 뉴럴 네트워크가 상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성정보를 학습하는 단계를 포함하는 음성인식 학습장치의 동작방법.

**청구항 15**

영상 입력부가 입력영상을 수신하는 단계;

영상 처리부가 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나누는 단계; 및

컨볼루션 뉴럴 네트워크가 상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성인식 결과를 결정하는 단계를 포함하는 음성인식 장치의 동작방법.

**청구항 16**

입력영상을 수신하는 영상 입력부;

미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나누는 영상 처리부; 및

상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성정보를 학습하는 컨볼루션 뉴럴 네트워크를 포함하는 음성인식 학습장치.

**청구항 17**

입력영상을 수신하는 영상 입력부;

미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나누는 영상 처리부; 및

상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성인식 결과를 결정하는 컨볼루션 뉴럴 네트워크를 포함하는 음성인식 장치.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 발명은 음성인식 학습장치, 음성인식 장치, 음성인식 학습장치의 동작방법 및 음성인식 장치의 동작방법에 관한 것이다.

### *배 경 기 술*

[0002] 마이크를 통해서 입력되는 소리 입력신호는 음성인식에 필요한 타겟 음성뿐만 아니라 음성인식에 방해가 되는 노이즈들이 포함될 수 있다. 소리 입력신호에서 노이즈를 제거하고, 원하는 타겟 음성만을 추출하여 음성인식의 성능을 높이기 위하여 화자를 포함하는 영상들을 활용하기도 한다. 최근, 이와 관련하여 다양한 연구가 진행되고 있다.

### *선행기술문헌*

#### *특허문헌*

[0003] (특허문헌 0001) (한국공개특허) 제10-2019-0016733호 (공개일자, 2019.02.19)

## *발명의 내용*

### *해결하려는 과제*

[0004] 본 발명이 이루고자 하는 기술적 과제는 화자의 입술영상 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시키는 음성인식 장치를 제공하는 것이다.

[0005] 본 발명이 이루고자 하는 기술적 과제는 화자의 입술영상 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상을 이용하여 컨볼루션 뉴럴 네트워크를 학습시킴으로써 음성인식 성능을 향상시키는

음성인식 장치의 동작방법을 제공하는 것이다.

## 과제의 해결 수단

[0006] 이러한 과제를 해결하기 위하여 본 발명의 실시예에 따른 음성인식 학습장치는 영상 입력부, 영상 처리부 및 컨볼루션 뉴럴 네트워크를 포함할 수 있다. 영상 입력부는 입력영상을 수신할 수 있다. 영상 처리부는 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크는 상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성정보를 학습할 수 있다.

[0007] 일 실시예에 있어서, 상기 영상 처리부는 라벨 변형기를 포함할 수 있다. 라벨 변형기는 상기 제1 시간간격 동안의 상기 입술영상 및 상기 패치영상의 길이를 조절하여 N개의 프레임들로 나눌 수 있다.

[0008] 일 실시예에 있어서, 상기 라벨 변형기는 영상길이 조절기 및 프레임 영상 제공기를 포함할 수 있다. 영상길이 조절기는 상기 제1 시간간격 동안의 상기 입술영상 및 상기 패치영상의 길이를 조절하여 길이조절 입술영상 및 길이조절 패치영상을 제공할 수 있다. 프레임 영상 제공기는 상기 길이조절 입술영상 및 상기 길이조절 패치영상을 각각 N개의 프레임들로 나누어 상기 프레임 입술영상 및 상기 프레임 패치영상을 제공할 수 있다.

[0009] 일 실시예에 있어서, 상기 입술영상의 길이를 조절하여 상기 입술영상의 길이가 증가하는 경우, 상기 패치영상의 길이를 조절하여 상기 패치영상의 길이가 증가될 수 있다.

[0010] 일 실시예에 있어서, 상기 입술영상의 길이를 조절하여 상기 입술영상의 길이가 감소하는 경우, 상기 패치영상의 길이를 조절하여 상기 패치영상의 길이가 감소될 수 있다.

[0011] 일 실시예에 있어서, 상기 컨볼루션 뉴럴 네트워크는 제1 컨볼루션 뉴럴 네트워크 및 제2 컨볼루션 뉴럴 네트워크를 포함할 수 있다.

[0012] 일 실시예에 있어서, 상기 제1 컨볼루션 뉴럴 네트워크는 상기 프레임 입술영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 학습할 수 있다.

[0013] 일 실시예에 있어서, 상기 제2 컨볼루션 뉴럴 네트워크는 상기 프레임 패치영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 학습할 수 있다.

[0014] 이러한 과제를 해결하기 위하여 본 발명의 실시예에 따른 음성인식 장치는 영상 입력부, 영상 처리부 및 컨볼루션 뉴럴 네트워크를 포함할 수 있다. 영상 입력부는 입력영상을 수신할 수 있다. 영상 처리부는 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크는 상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성인식 결과를 결정할 수 있다.

[0015] 일 실시예에 있어서, 상기 컨볼루션 뉴럴 네트워크는 제1 컨볼루션 뉴럴 네트워크 및 제2 컨볼루션 뉴럴 네트워크를 포함할 수 있다.

[0016] 일 실시예에 있어서, 상기 제1 컨볼루션 뉴럴 네트워크는 상기 프레임 입술영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 결정할 수 있다.

[0017] 일 실시예에 있어서, 상기 제2 컨볼루션 뉴럴 네트워크는 상기 프레임 패치영상에 기초하여 상기 입술 모양에 상응하는 음성정보를 결정할 수 있다.

[0018] 일 실시예에 있어서, 상기 제1 컨볼루션 뉴럴 네트워크로부터 결정되는 음성정보에 상응하는 제1 확률 및 상기 제2 컨볼루션 뉴럴 네트워크로부터 결정되는 음성정보에 상응하는 제2 확률의 합에 기초하여 음성인식 결과가 결정될 수 있다.

[0019] 이러한 과제를 해결하기 위하여 본 발명의 실시예에 따른 음성인식 학습장치의 동작방법에서는, 영상 입력부가 입력영상을 수신할 수 있다. 영상 처리부가 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크가 상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성정보를

학습할 수 있다.

[0020] 이러한 과제를 해결하기 위하여 본 발명의 실시예에 따른 음성인식 장치의 동작방법에서는, 영상 입력부가 입력영상을 수신할 수 있다. 영상 처리부가 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술영상 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크가 상기 입술영상을 N개의 프레임들로 나눈 프레임 입술영상 및 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성인식 결과를 결정할 수 있다.

[0021] 이러한 과제를 해결하기 위하여 본 발명의 실시예들에 따른 음성인식 학습장치는 영상 입력부, 영상 처리부 및 컨볼루션 뉴럴 네트워크를 포함할 수 있다. 영상 입력부는 입력영상을 수신할 수 있다. 영상 처리부는 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크는 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성정보를 학습할 수 있다.

[0022] 이러한 과제를 해결하기 위하여 본 발명의 실시예들에 따른 음성인식 장치는 영상 입력부, 영상 처리부 및 컨볼루션 뉴럴 네트워크를 포함할 수 있다. 영상 입력부는 입력영상을 수신할 수 있다. 영상 처리부는 미리 정해진 제1 시간간격 동안의 상기 입력영상에 포함되는 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크는 상기 패치영상을 N개의 프레임들로 나눈 프레임 패치영상에 기초하여 상기 입술의 모양에 상응하는 음성인식 결과를 결정할 수 있다.

[0023] 위에서 언급된 본 발명의 기술적 과제 외에도, 본 발명의 다른 특징 및 이점들이 이하에서 기술되거나, 그러한 기술 및 설명으로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

## *발명의 효과*

[0024] 이상과 같은 본 발명에 따르면 다음과 같은 효과가 있다.

[0025] 본 발명에 따른 음성인식 장치에서는 화자의 입술영상 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

[0026] 본 발명에 따른 음성인식 장치의 동작방법에서는 화자의 입술영상 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

[0027] 이 밖에도, 본 발명의 실시 예들을 통해 본 발명의 또 다른 특징 및 이점들이 새롭게 파악될 수도 있을 것이다.

## *도면의 간단한 설명*

[0028] 도 1은 본 발명의 실시예들에 따른 음성인식 학습장치를 나타내는 도면이다.

도 2는 도 1의 음성인식 학습장치에 포함되는 영상 처리부에서 사용되는 입술영상의 일 예를 나타내는 도면이다.

도 3은 도 1의 음성인식 학습장치에 포함되는 영상 처리부에서 사용되는 패치영상의 일 예를 나타내는 도면이다.

도 4는 도 1의 음성인식 학습장치에 포함되는 영상 처리부로부터 출력되는 프레임 입술영상의 일 예를 나타내는 도면이다.

도 5는 도 1의 음성인식 학습장치에 포함되는 영상 처리부로부터 출력되는 프레임 패치영상의 일 예를 나타내는 도면이다.

도 6은 도 1의 음성인식 학습장치에 포함되는 영상 처리부의 일 예를 나타내는 도면이다.

도 7은 도 6의 영상 처리부에 포함되는 라벨 변형기의 일 예를 나타내는 도면이다.

도 8 및 9는 도 6의 영상 처리부에 포함되는 라벨 변형기의 동작을 설명하기 위한 도면들이다.

도 10은 도 1의 음성인식 학습장치에 포함되는 컨볼루션 뉴럴 네트워크의 일 예를 나타내는 도면이다.

도 11은 본 발명의 실시예들에 따른 음성인식 장치를 나타내는 도면이다.

도 12는 도 11의 음성인식 장치에 포함되는 컨볼루션 뉴럴 네트워크의 일 예를 나타내는 도면이다.

도 13은 본 발명의 실시예들에 따른 음성인식 학습장치의 동작방법을 나타내는 도면이다.

도 14는 본 발명의 실시예들에 따른 음성인식 장치의 동작방법을 나타내는 도면이다.

도 15는 본 발명의 실시예들에 따른 음성인식 학습장치를 나타내는 도면이다.

도 16은 본 발명의 실시예들에 따른 음성인식 장치를 나타내는 도면이다.

## *발명을 실시하기 위한 구체적인 내용*

[0029] 본 명세서에서 각 도면의 구성 요소들에 참조번호를 부가함에 있어서 동일한 구성 요소들에 한해서는 비록 다른 도면상에 표시되더라도 가능한한 동일한 번호를 가지도록 하고 있음에 유의하여야 한다.

[0030] 한편, 본 명세서에서 서술되는 용어의 의미는 다음과 같이 이해되어야 할 것이다.

[0031] 단수의 표현은 문맥상 명백하게 다르게 정의하지 않는 한, 복수의 표현을 포함하는 것으로 이해되어야 하는 것으로, 이들 용어들에 의해 권리범위가 한정되어서는 아니 된다.

[0032] "포함하다" 또는 "가지다" 등의 용어는 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0033] 이하, 첨부되는 도면을 참고하여 상기 문제점을 해결하기 위해 고안된 본 발명의 바람직한 실시예들에 대해 상세히 설명한다.

[0034] 도 1은 본 발명의 실시예들에 따른 음성인식 학습장치를 나타내는 도면이고, 도 2는 도 1의 음성인식 학습장치에 포함되는 영상 처리부에서 사용되는 입술영상의 일 예를 나타내는 도면이고, 도 3은 도 1의 음성인식 학습장치에 포함되는 영상 처리부에서 사용되는 패치영상의 일 예를 나타내는 도면이고, 도 4는 도 1의 음성인식 학습장치에 포함되는 영상 처리부로부터 출력되는 프레임 입술영상의 일 예를 나타내는 도면이고, 도 5는 도 1의 음성인식 학습장치에 포함되는 영상 처리부로부터 출력되는 프레임 패치영상의 일 예를 나타내는 도면이다.

[0035] 도 1 내지 5를 참조하면, 본 발명의 실시예에 따른 음성인식 학습장치(10)는 영상 입력부(100), 영상 처리부(200) 및 컨볼루션 뉴럴 네트워크(300)를 포함할 수 있다. 영상 입력부(100)는 입력영상(IM_IN)을 수신할 수 있다. 입력영상(IM_IN)은 입술영상(RI) 및 패치영상(PI)을 포함할 수 있다. 입술영상(RI)은 입력영상(IM_IN)에서 화자의 입술부분을 포함한 영상일 수 있다. 패치영상(PI)은 입술의 정해진 위치에 배치되는 랜드마크 주변의 일정영역에 대한 영상일 수 있다. 예를 들어, 랜드마크(LM)는 제1 랜드마크(LM1) 내지 제5 랜드마크(LM5)를 포함할 수 있다. 제1 랜드마크(LM1) 주변의 일정영역에 대한 영상은 제1 패치영상(PI1)일 수 있고, 제2 랜드마크(LM2) 주변의 일정영역에 대한 영상은 제2 패치영상(PI2)일 수 있다. 동일한 방식으로, 제5 랜드마크(LM5) 주변의 일정영역에 대한 영상은 제5 패치영상(PI5)일 수 있다.

[0036] 영상 처리부(200)는 미리 정해진 제1 시간간격(TP1) 동안의 입력영상(IM_IN)에 포함되는 입술영상(RI) 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상(PI)을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다.

[0037] 예를 들어, 제1 시간간격(TP1)은 제1 시간(T1)부터 제N 시간(TN) 사이의 시간 간격일 수 있다. 영상 처리부(200)는 제1 시간간격(TP1) 동안의 입술영상(RI)을 프레임 간격(FD)으로 나누어 N개의 프레임 입술영상(F_RI)을 제공할 수 있다. 이 경우, 프레임 입술영상(F_RI)은 제1 프레임 입술영상(F_RI1) 내지 제N 프레임 입술영상(F_RIN)을 포함할 수 있다. 또한, 영상 처리부(200)는 제1 시간간격(TP1) 동안의 제1 패치영상(PI1)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다. 이 경우, 프레임 패치영상(F_PI)은 제1_1 프레임 패치영상(F_PI1_1) 내지 제1_N 프레임 패치영상(F_PI1_N)을 포함할 수 있다. 또한, 영상 처리부(200)는 제1 시간간격(TP1) 동안의 제2 패치영상(PI2)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다. 이 경우, 프레임 패치영상(F_PI)은 제2_1 프레임 패치영상(F_PI2_1) 내지 제2_N 프레임 패치영상(F_PI2_N)을 포함할 수 있다. 동일한 방식으로, 영상 처리부(200)는 제1 시간간격(TP1) 동안의 제K(K는 자연수)

패치영상(PI)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다. 이 경우, 프레임 패치영상(F_PI)은 제K_1 프레임 패치영상(F_PIK_1) 내지 제K_N 프레임 패치영상(F_PIK_N)을 포함할 수 있다.

[0038] 컨볼루션 뉴럴 네트워크(300)는 입술영상(RI)을 N개의 프레임들로 나눈 프레임 입술영상(F_RI) 및 패치영상(PI)을 N개의 프레임들로 나눈 프레임 패치영상(F_PI)에 기초하여 입술의 모양에 상응하는 음성정보를 학습할 수 있다.

[0039] 본 발명에 따른 음성인식 장치에서는 화자의 입술영상(RI) 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상(PI)을 이용하여 컨볼루션 뉴럴 네트워크(300)(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

[0040] 도 6은 도 1의 음성인식 학습장치에 포함되는 영상 처리부의 일 예를 나타내는 도면이고, 도 7은 도 6의 영상 처리부에 포함되는 라벨 변형기의 일 예를 나타내는 도면이고, 도 8 및 9는 도 6의 영상 처리부에 포함되는 라벨 변형기의 동작을 설명하기 위한 도면들이다.

[0041] 도 6 내지 9를 참조하면, 영상 처리부(200)는 라벨 변형기(210)를 포함할 수 있다. 라벨 변형기(210)는 제1 시간간격(TP1) 동안의 입술영상(RI) 및 패치영상(PI)의 길이를 조절하여 N개의 프레임들로 나눌 수 있다. 입술영상(RI) 및 패치영상(PI)의 길이를 조절하여 N개의 프레임들로 나누는 경우, 입술영상(RI) 및 패치영상(PI)을 조절하는 길이에 따라 다양한 프레임 입술영상(F_RI) 및 프레임 패치영상(F_PI)을 구성할 수 있다. 이 경우, 제1 컨볼루션 뉴럴 네트워크(310) 및 제2 컨볼루션 뉴럴 네트워크(320)에 제공되는 프레임 입술영상(F_RI) 및 프레임 패치영상(F_PI)의 데이터 양이 증가할 수 있다. 제1 컨볼루션 뉴럴 네트워크(310) 및 제2 컨볼루션 뉴럴 네트워크(320)에 제공되는 프레임 입술영상(F_RI) 및 프레임 패치영상(F_PI)의 데이터 양이 증가하는 경우, 본 발명에 따른 음성인식 학습장치 및 음성인식 장치의 성능은 향상될 수 있다.

[0042] 예를 들어, 라벨 변형기(210)는 일정시간(A1) 동안의 앞부분 입술영상(RI) 및 뒷부분 입술영상(RI)을 연장하여 입술영상(RI)의 길이를 증가시킬 수 있고, 또한, 라벨 변형기(210)는 일정시간(A1) 동안의 앞부분 입술영상(RI) 및 뒷부분 입술영상(RI)을 삭제하여 입술영상(RI)의 길이를 감소시킬 수 있다.

[0043] 예를들어, 라벨 변형기(210)는 일정시간(A1) 동안의 앞부분 패치영상(PI) 및 뒷부분 패치영상(PI)을 연장하여 패치영상(PI)의 길이를 증가시킬 수 있고, 또한, 라벨 변형기(210)는 일정시간(A1) 동안의 앞부분 패치영상(PI) 및 뒷부분 패치영상(PI)을 삭제하여 패치영상(PI)의 길이를 감소시킬 수 있다.

[0044] 일 실시예에 있어서, 라벨 변형기(210)는 영상길이 조절기(211) 및 프레임 영상 제공기(213)를 포함할 수 있다. 영상길이 조절기(211)는 제1 시간간격(TP1) 동안의 입술영상(RI) 및 패치영상(PI)의 길이를 조절하여 길이조절 입술영상(LC_RI) 및 길이조절 패치영상(LC_PI)을 제공할 수 있다. 프레임 영상 제공기(213)는 길이조절 입술영상(LC_RI) 및 길이조절 패치영상(LC_PI)을 각각 N개의 프레임들로 나누어 프레임 입술영상(F_RI) 및 프레임 패치영상(F_PI)을 제공할 수 있다.

[0045] 예를 들어, 영상길이 조절기(211)는 일정시간(A1) 동안의 앞부분 입술영상(RI) 및 뒷부분 입술영상(RI)을 연장하여 입술영상(RI)의 길이를 제2 시간간격(TP2)으로 증가시킬 수 있다. 이 경우, 제2 시간간격(TP2)의 입술영상(RI)은 길이조절 입술영상(LC_RI)일 수 있다. 프레임 영상 제공기(213)는 제2 시간간격(TP2)의 입술영상(RI)을 프레임 간격(FD)으로 나누어 N개의 프레임 입술영상(F_RI)을 제공할 수 있다. 또한, 영상길이 조절기(211)는 일정시간(A1) 동안의 앞부분 패치영상(PI) 및 뒷부분 패치영상(PI)을 연장하여 패치영상(PI)의 길이를 제2 시간간격(TP2)으로 증가시킬 수 있다. 이 경우, 제2 시간간격(TP2)의 패치영상(PI)은 길이조절 패치영상(LC_PI)일 수 있다. 프레임 영상 제공기(213)는 제2 시간간격(TP2)의 패치영상(PI)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다.

[0046] 예를 들어, 영상길이 조절기(211)는 일정시간(A1) 동안의 앞부분 입술영상(RI) 및 뒷부분 입술영상(RI)을 삭제하여 입술영상(RI)의 길이를 제3 시간간격(TP3)으로 감소시킬 수 있다. 이 경우, 제3 시간간격(TP3)의 입술영상(RI)은 길이조절 입술영상(LC_RI)일 수 있다. 프레임 영상 제공기(213)는 제3 시간간격(TP3)의 입술영상(RI)을 프레임 간격(FD)으로 나누어 N개의 프레임 입술영상(F_RI)을 제공할 수 있다. 또한, 영상길이 조절기(211)는 일정시간(A1) 동안의 앞부분 패치영상(PI) 및 뒷부분 패치영상(PI)을 삭제하여 패치영상(PI)의 길이를 제3 시간간격(TP3)으로 감소시킬 수 있다. 이 경우, 제3 시간간격(TP3)의 패치영상(PI)은 길이조절 패치영상(LC_PI)일 수 있다. 프레임 영상 제공기(213)는 제3 시간간격(TP3)의 패치영상(PI)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다.

[0047] 일 실시예에 있어서, 입술영상(RI)의 길이를 조절하여 입술영상(RI)의 길이가 증가하는 경우, 패치영상(PI)의

길이를 조절하여 패치영상(PI)의 길이가 증가될 수 있다.

[0048] 일 실시예에 있어서, 입술영상(RI)의 길이를 조절하여 입술영상(RI)의 길이가 감소하는 경우, 패치영상(PI)의 길이를 조절하여 패치영상(PI)의 길이가 감소될 수 있다.

[0049] 도 10은 도 1의 음성인식 학습장치에 포함되는 컨볼루션 뉴럴 네트워크의 일 예를 나타내는 도면이다.

[0050] 도 10을 참조하면, 컨볼루션 뉴럴 네트워크(300)는 제1 컨볼루션 뉴럴 네트워크(310) 및 제2 컨볼루션 뉴럴 네트워크(320)를 포함할 수 있다. 일 실시예에 있어서, 제1 컨볼루션 뉴럴 네트워크(310)는 프레임 입술영상(F_RI)에 기초하여 입술 모양에 상응하는 음성정보(SRR1)를 학습할 수 있다. 일 실시예에 있어서, 제2 컨볼루션 뉴럴 네트워크(320)는 프레임 패치영상(F_PI)에 기초하여 입술 모양에 상응하는 음성정보(SRR2)를 학습할 수 있다.

[0051] 또한, 제1 컨볼류션 뉴럴 네트워크(310)는 제1_1 컨볼루션 뉴럴 네트워크 내지 제1_N 컨볼루션 뉴럴 네트워크(N은 2 이상의 자연수)를 포함할 수 있다. 제1_1 컨볼루션 뉴럴 네트워크 내지 제1_N 컨볼루션 뉴럴 네트워크를 이용해 프레임 입술영상(F_RI)에 기초하여 입술 모양에 상응하는 음성정보(SRR1)를 학습하는 경우, 본 발명에 따른 음성인식 학습장치의 성능을 더욱 향상시킬 수 있다. 또한, 제2 컨볼류션 뉴럴 네트워크(320)는 제2_1 컨볼루션 뉴럴 네트워크 내지 제2_N 컨볼루션 뉴럴 네트워크(N은 2 이상의 자연수)를 포함할 수 있다. 제2_1 컨볼루션 뉴럴 네트워크 내지 제2_N 컨볼루션 뉴럴 네트워크를 이용해 프레임 패치영상(F_RI)에 기초하여 입술 모양에 상응하는 음성정보(SRR2)를 학습하는 경우, 본 발명에 따른 음성인식 학습장치의 성능을 더욱 향상시킬 수 있다.

[0052] 본 발명에 따른 음성인식 장치에서는 화자의 입술영상(RI) 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상(PI)을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

[0053] 도 11은 본 발명의 실시예들에 따른 음성인식 장치를 나타내는 도면이고, 도 12는 도 11의 음성인식 장치에 포함되는 컨볼루션 뉴럴 네트워크의 일 예를 나타내는 도면이다.

[0054] 도 1 내지 5, 11 및 12를 참조하면, 본 발명의 실시예에 따른 음성인식 장치는 영상 입력부(100), 영상 처리부(200) 및 컨볼루션 뉴럴 네트워크(300)를 포함할 수 있다. 영상 입력부(100)는 입력영상(IM_IN)을 수신할 수 있다. 입력영상(IM_IN)은 입술영상(RI) 및 패치영상(PI)을 포함할 수 있다. 입술영상(RI)은 입력영상(IM_IN)에서 화자의 입술부분을 포함한 영상일 수 있다. 패치영상(PI)은 입술의 정해진 위치에 배치되는 랜드마크의 주변의 일정영역에 대한 영상일 수 있다. 예를 들어, 랜드마크는 제1 랜드마크(LM1) 내지 제5 랜드마크(LM5)를 포함할 수 있다. 제1 랜드마크(LM1) 주변의 일정영역에 대한 영상은 제1 패치영상(PI1)일 수 있고, 제2 랜드마크(LM2) 주변의 일정영역에 대한 영상은 제2 패치영상(PI2)일 수 있다. 동일한 방식으로, 제5 랜드마크(LM5) 주변의 일정영역에 대한 영상은 제5 패치영상(PI5)일 수 있다.

[0055] 영상 처리부(200)는 미리 정해진 제1 시간간격(TP1) 동안의 입력영상(IM_IN)에 포함되는 입술영상(RI) 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상(PI)을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다.

[0056] 예를 들어, 제1 시간간격(TP1)은 제1 시간(T1)부터 제N 시간(TN) 사이의 시간 간격일 수 있다. 영상 처리부(200)는 제1 시간간격(TP1) 동안의 입술영상(RI)을 프레임 간격(FD)으로 나누어 N개의 프레임 입술영상(F_RI)을 제공할 수 있다. 이 경우, 프레임 입술영상(F_RI)은 제1 프레임 입술영상(F_RI1) 내지 제N 프레임 입술영상(F_RIN)을 포함할 수 있다. 또한, 영상 처리부(200)는 제1 시간간격(TP1) 동안의 제1 패치영상(PI1)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다. 이 경우, 프레임 패치영상(F_PI)은 제1_1 프레임 패치영상(F_PI1_1) 내지 제1_N 프레임 패치영상(F_PI1_N)을 포함할 수 있다. 또한, 영상 처리부(200)는 제1 시간간격(TP1) 동안의 제2 패치영상(PI2)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다. 이 경우, 프레임 패치영상(F_PI)은 제2_1 프레임 패치영상(F_PI2_1) 내지 제2_N 프레임 패치영상(F_PI2_N)을 포함할 수 있다. 동일한 방식으로, 영상 처리부(200)는 제1 시간간격(TP1) 동안의 제K(K는 자연수) 패치영상(PI)을 프레임 간격(FD)으로 나누어 N개의 프레임 패치영상(F_PI)을 제공할 수 있다. 이 경우, 프레임 패치영상(F_PI)은 제K_1 프레임 패치영상(F_PIK_1) 내지 제K_N 프레임 패치영상(F_PIK_N)을 포함할 수 있다.

[0057] 컨볼루션 뉴럴 네트워크(300)는 입술영상(RI)을 N개의 프레임들로 나눈 프레임 입술영상(F_RI) 및 패치영상(PI)을 N개의 프레임들로 나눈 프레임 패치영상(F_PI)에 기초하여 입술의 모양에 상응하는 음성인식 결과(SRR)를 결정할 수 있다. 예를 들어, 컨볼루션 뉴럴 네트워크(300)는 제1 컨볼루션 뉴럴 네트워크(310) 및 제2 컨볼루션

뉴럴 네트워크(320)를 포함할 수 있다. 일 실시예에 있어서, 제1 컨볼루션 뉴럴 네트워크(310)는 프레임 입술영상(F_RI)에 기초하여 입술 모양에 상응하는 음성정보(SRR1)를 결정할 수 있고, 제2 컨볼루션 뉴럴 네트워크(320)는 프레임 패치영상(F_PI)에 기초하여 입술 모양에 상응하는 음성정보(SRR2)를 결정할 수 있다.

[0058] 또한, 제1 컨볼류션 뉴럴 네트워크(310)는 제1_1 컨볼루션 뉴럴 네트워크 내지 제1_N 컨볼루션 뉴럴 네트워크(N은 2 이상의 자연수)를 포함할 수 있다. 제1_1 컨볼루션 뉴럴 네트워크 내지 제1_N 컨볼루션 뉴럴 네트워크를 이용해 프레임 입술영상(F_RI)에 기초하여 입술 모양에 상응하는 음성정보(SRR1)를 결정하는 경우, 본 발명에 따른 음성인식 장치의 성능을 더욱 향상시킬 수 있다. 또한, 제2 컨볼류션 뉴럴 네트워크(320)는 제2_1 컨볼루션 뉴럴 네트워크 내지 제2_N 컨볼루션 뉴럴 네트워크(N은 2 이상의 자연수)를 포함할 수 있다. 제2_1 컨볼루션 뉴럴 네트워크 내지 제2_N 컨볼루션 뉴럴 네트워크를 이용해 프레임 패치영상(F_RI)에 기초하여 입술 모양에 상응하는 음성정보(SRR2)를 결정하는 경우, 본 발명에 따른 음성인식 장치의 성능을 더욱 향상시킬 수 있다.

[0059] 일 실시예에 있어서, 제1 컨볼루션 뉴럴 네트워크(310)로부터 결정되는 음성정보에 상응하는 제1 확률(P1) 및 제2 컨볼루션 뉴럴 네트워크(320)로부터 결정되는 음성정보에 상응하는 제2 확률(P2)의 합에 기초하여 음성인식 결과(SRR)가 결정될 수 있다. 예를 들어, 제1 컨볼루션 뉴럴 네트워크(310)는 입술영상(RI)에 상응하는 음성정보를 제1 문장 또는 제2 문장으로 판단할 수 있다. 이 경우, 제1 컨볼루션 뉴럴 네트워크(310)는 제1 문장 또는 제2 문장 각각이 정답일 확률을 제1 확률(P1)로서 제공할 수 있다. 예를 들어, 제1 컨볼루션 뉴럴 네트워크(310)로부터 제공되는 제1 문장이 정답일 제1 확률(P1)은 0.8일 수 있고, 제1 컨볼루션 뉴럴 네트워크(310)로부터 제공되는 제2 문장이 정답일 제1 확률(P1)은 0.7일 수 있다.

[0060] 또한, 제2 컨볼루션 뉴럴 네트워크(320)는 패치영상(PI)에 상응하는 음성정보를 제1 문장 또는 제2 문장으로 판단할 수 있다. 이 경우, 제2 컨볼루션 뉴럴 네트워크(320)는 제1 문장 또는 제2 문장 각각이 정답일 확률을 제2 확률(P2)로서 제공할 수 있다. 예를 들어, 제2 컨볼루션 뉴럴 네트워크(320)로부터 제공되는 제1 문장이 정답일 제2 확률(P2)은 0.3일 수 있고, 제2 컨볼루션 뉴럴 네트워크(320)로부터 제공되는 제2 문장이 정답일 제2 확률(P2)은 0.7일 수 있다.

[0061] 이 경우, 제1 컨볼루션 뉴럴 네트워크(310)로부터 결정되는 음성정보에 상응하는 제1 확률(P1) 및 제2 컨볼루션 뉴럴 네트워크(320)로부터 결정되는 음성정보에 상응하는 제2 확률(P2)의 합에 기초하여 음성인식 결과(SRR)가 결정될 수 있다. 예를 들어, 제1 문장의 경우, 제1 확률(P1) 및 제2 확률(P2)의 합은 1.1일 수 있고, 제2 문장의 경우, 제1 확률(P1) 및 제2 확률(P2)의 합은 1.4일 수 있다. 이 경우, 영상입력에 상응하는 음성인식 결과(SRR)는 제2 문장일 수 있다.

[0062] 본 발명에 따른 음성인식 장치에서는 화자의 입술영상(RI) 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상(PI)을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

[0063] 도 13은 본 발명의 실시예들에 따른 음성인식 학습장치의 동작방법을 나타내는 도면이다.

[0064] 도 13을 참조하면, 본 발명의 실시예에 따른 음성인식 학습장치(10)의 동작방법에서는, 영상 입력부(100)가 입력영상(IM_IN)을 수신할 수 있다(S100). 영상 처리부(200)가 미리 정해진 제1 시간간격(TP1) 동안의 입력영상(IM_IN)에 포함되는 입술영상(RI) 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상(PI)을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다(S200). 컨볼루션 뉴럴 네트워크(300)가 입술영상(RI)을 N개의 프레임들로 나눈 프레임 입술영상(F_RI) 및 패치영상(PI)을 N개의 프레임들로 나눈 프레임 패치영상(F_PI)에 기초하여 입술의 모양에 상응하는 음성정보를 학습할 수 있다(S300).

[0065] 도 14는 본 발명의 실시예들에 따른 음성인식 장치의 동작방법을 나타내는 도면이다.

[0066] 도 14를 참조하면, 본 발명의 실시예에 따른 음성인식 장치의 동작방법에서는, 영상 입력부(100)가 입력영상(IM_IN)을 수신할 수 있다(S400). 영상 처리부(200)가 미리 정해진 제1 시간간격(TP1) 동안의 입력영상(IM_IN)에 포함되는 입술영상(RI) 및 입술의 정해진 위치에 상응하는 랜드마크의 주변영상에 해당하는 패치영상(PI)을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다(S500). 컨볼루션 뉴럴 네트워크(300)가 입술영상(RI)을 N개의 프레임들로 나눈 프레임 입술영상(F_RI) 및 패치영상(PI)을 N개의 프레임들로 나눈 프레임 패치영상(F_PI)에 기초하여 입술의 모양에 상응하는 음성인식 결과(SRR)를 결정할 수 있다(S600).

[0067] 본 발명에 따른 음성인식 장치에서는 화자의 입술영상(RI) 뿐만 아니라, 입술의 정해진 위치에 배치되는 랜드마크의 주변영상인 패치영상(PI)을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시

킴으로써 음성인식 성능을 향상시킬 수 있다.

[0068] 도 15는 본 발명의 실시예들에 따른 음성인식 학습장치를 나타내는 도면이다.

[0069] 도 15를 참조하면, 본 발명의 실시예들에 따른 음성인식 학습장치(10)는 영상 입력부(100), 영상 처리부(200) 및 컨볼루션 뉴럴 네트워크(300)를 포함할 수 있다. 영상 입력부(100)는 입력영상(IM_IN)을 수신할 수 있다. 영상 처리부(200)는 미리 정해진 제1 시간간격(TP1) 동안의 입력영상(IM_IN)에 포함되는 입술의 정해진 위치에 상응하는 랜드마크(LM)의 주변영상에 해당하는 패치영상(PI)을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크(300)는 패치영상(PI)을 N개의 프레임들로 나눈 프레임 패치영상(F_PI)에 기초하여 입술의 모양에 상응하는 음성정보를 학습할 수 있다.

[0070] 도 16은 본 발명의 실시예들에 따른 음성인식 장치를 나타내는 도면이다.

[0071] 도 16을 참조하면, 본 발명의 실시예들에 따른 음성인식 장치는 영상 입력부(100), 영상 처리부(200) 및 컨볼루션 뉴럴 네트워크(300)를 포함할 수 있다. 영상 입력부(100)는 입력영상(IM_IN)을 수신할 수 있다. 영상 처리부(200)는 미리 정해진 제1 시간간격(TP1) 동안의 입력영상(IM_IN)에 포함되는 입술의 정해진 위치에 상응하는 랜드마크(LM)의 주변영상에 해당하는 패치영상(PI)을 각각 N(N은 자연수)개의 프레임들로 나눌 수 있다. 컨볼루션 뉴럴 네트워크(300)는 패치영상(PI)을 N개의 프레임들로 나눈 프레임 패치영상(F_PI)에 기초하여 입술의 모양에 상응하는 음성인식 결과(SRR)를 결정할 수 있다.

[0072] 본 발명에 따른 음성인식 장치에서는 입술의 정해진 위치에 배치되는 랜드마크(LM)의 주변영상인 패치영상(PI)을 이용하여 컨볼루션 뉴럴 네트워크(Convolution Neural Network, CNN)를 학습시킴으로써 음성인식 성능을 향상시킬 수 있다.

[0073] 위에서 언급된 본 발명의 기술적 과제 외에도, 본 발명의 다른 특징 및 이점들이 이하에서 기술되거나, 그러한 기술 및 설명으로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

## 부호의 설명

[0074] 10: 음성인식 학습장치 100: 영상 입력부

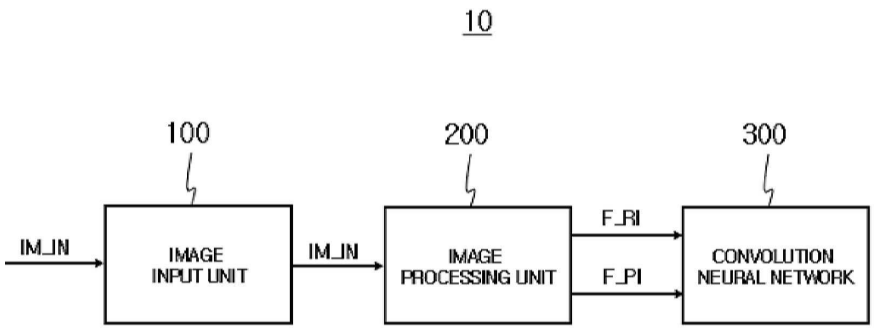
200: 영상 처리부 300: 컨볼루션 뉴럴 네트워크

210: 라벨 변형기 211: 영상길이 조절기

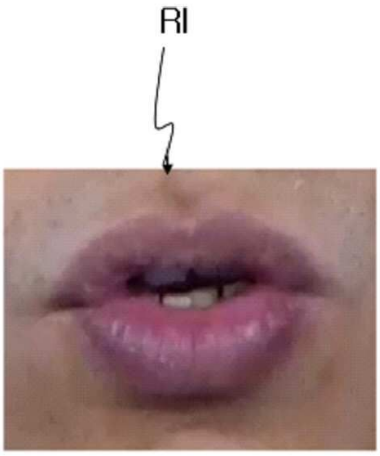
213: 프레임 영상 제공기

## 도면

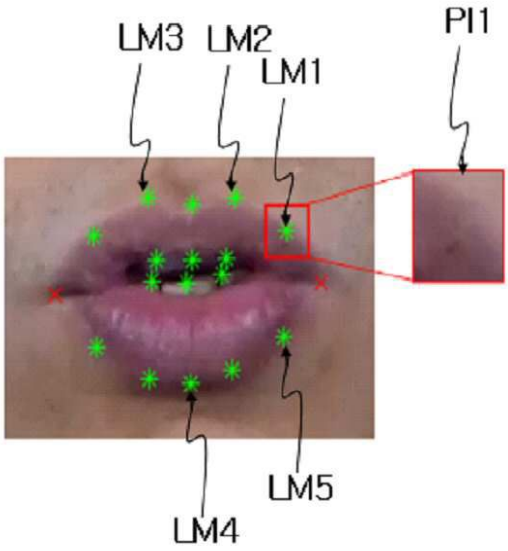
### 도면1



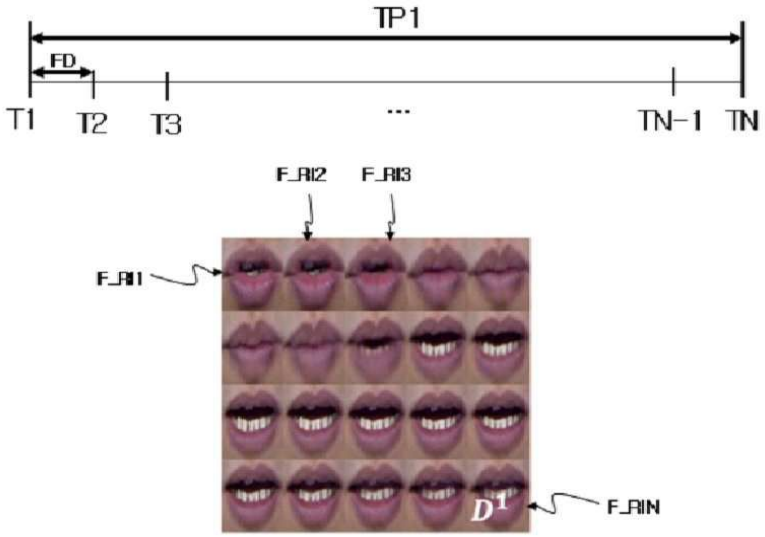
도면2



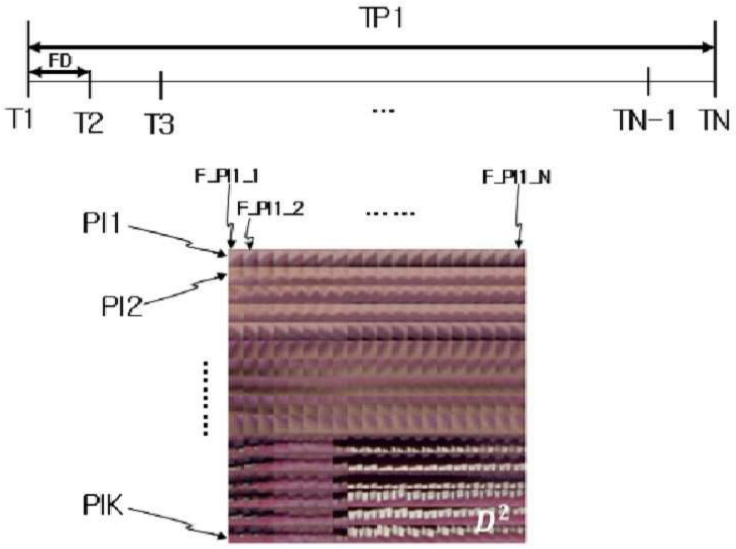

도면3



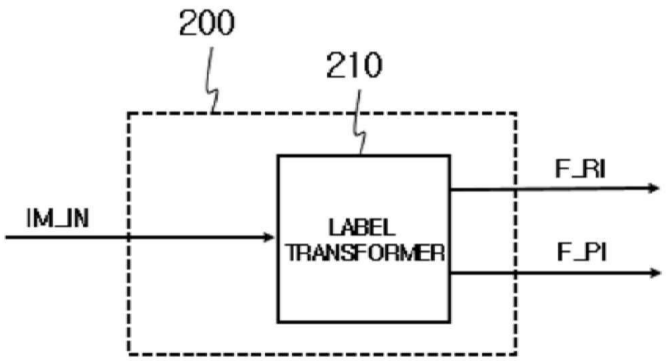
도면4


TP1
FD
T1
T2
T3
...
TN-1
TN
F_RI2
F_RI3
F_RI1
F_RIN
D1

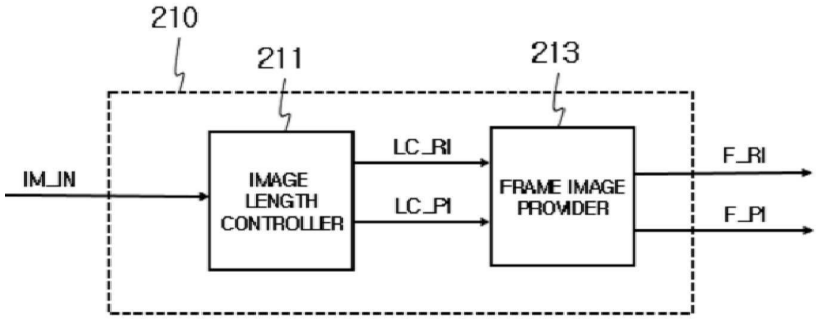
도면5


TP1
FD
T1
T2
T3
...
TN-1
TN
F_PI1_1
F_PI1_2
F_PI1_N
PI1
PI2
PIK
D2

도면6


200
210
IM_IN
LABEL
TRANSFORMER
F_RI
F_PI

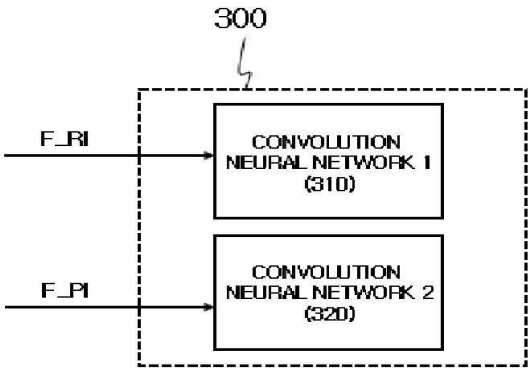
도면7



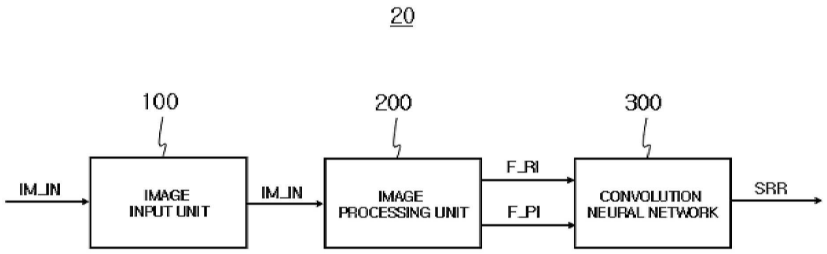

도면8


도면9

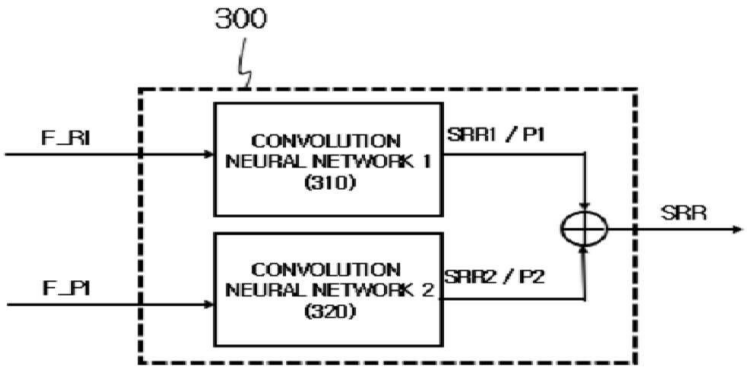

도면10



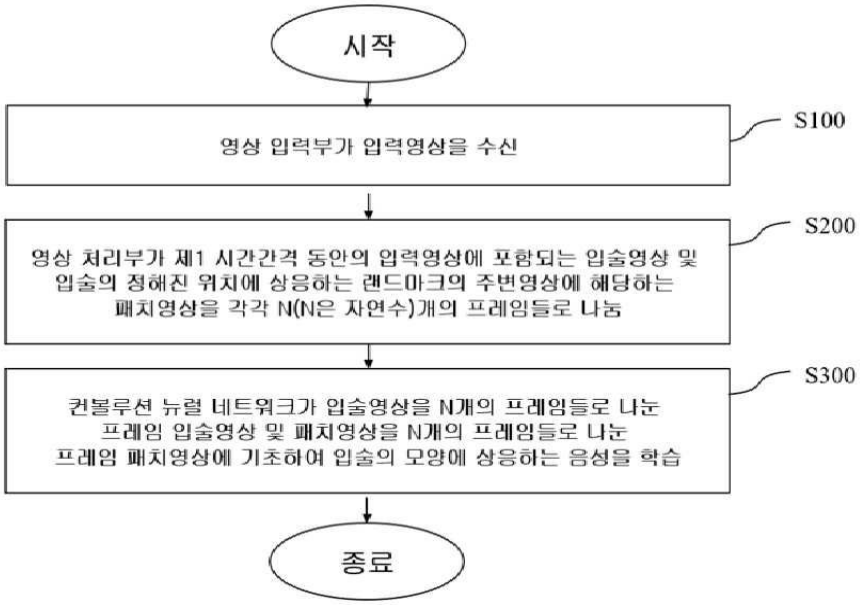
도면11



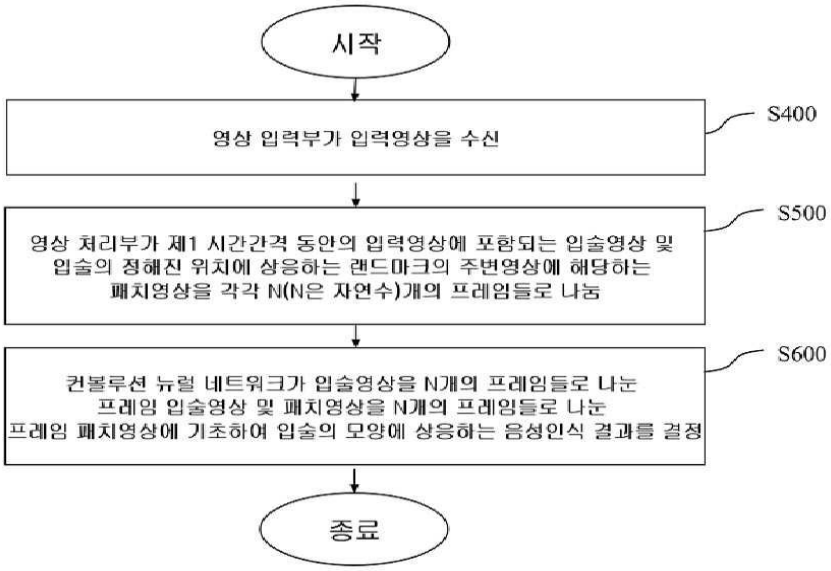

도면12



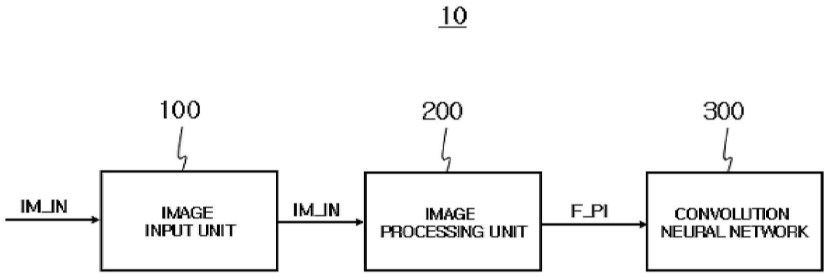

도면13



도면14




도면15




도면16