(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0075337
(43) 공개일자 2023년05월31일

(51) 국제특허분류(Int. Cl.)
G10L 15/00 (2006.01) *G10L 15/04* (2006.01)
G10L 15/16 (2006.01) *G10L 15/183* (2013.01)
G10L 15/22 (2006.01)
(52) CPC특허분류
G10L 15/005 (2013.01)
G10L 15/04 (2013.01)
(21) 출원번호 10-2022-0076334
(22) 출원일자 2022년06월22일
심사청구일자 2022년06월22일
(30) 우선권주장
1020210161358 2021년11월22일 대한민국(KR)

(71) 출원인
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
이근배
경상북도 포항시 남구 청암로 77
이원준
경상북도 포항시 남구 청암로 77
(74) 대리인
특허법인이상

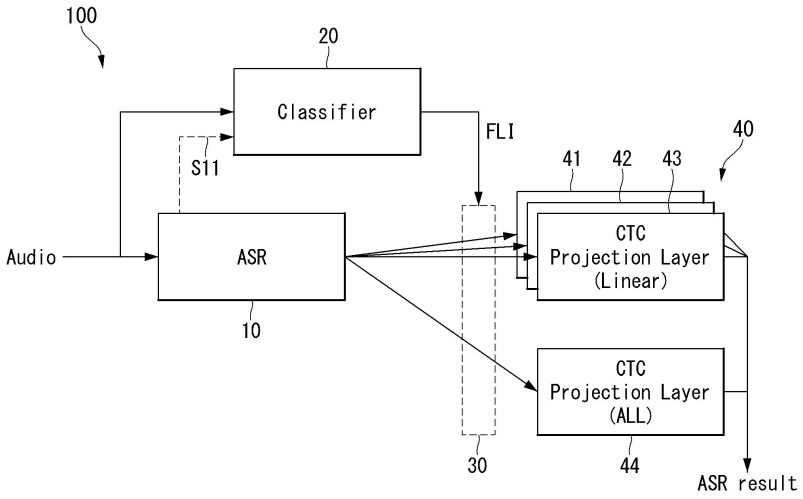
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **인공지능모델 기반 다국어 음성인식 방법 및 장치**

(57) 요 약

단일 모델의 인공지능 기반으로 다국어의 오디오 데이터를 자동으로 인식하는 다국어 자동음성인식 방법 및 장치가 개시된다. 다국어 자동음성인식 방법은, 음성인식기에 의해, 입력되는 오디오 데이터를 인식하는 단계, 음성언어분류기에 의해, 오디오 데이터를 분류하는 단계, 음성인식기에 결합된 출력계층선택기에 의해, 음성언어분류기로부터 받은 언어 분류 정보에 따라 음성인식기에 각각 연결된 복수의 프로젝션 출력 계층들 중 어느 하나의 프로젝션 출력 계층을 활성화하는 단계, 여기서 활성화된 프로젝션 출력 계층의 출력 단위는 1바이트로 설정되고, 그리고 활성화된 프로젝션 출력 계층에 의해, 1바이트의 출력 단위로 출력되는 출력들을 재조합하여 오디오 데이터에 대한 자동음성인식 결과로서 출력하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G10L 15/16 (2013.01)
G10L 15/183 (2013.01)
G10L 2015/221 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1415174008
과제번호 20015007
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 바이오산업기술개발
연구과제명 공황장애 환자 대상 인지행동치료 디지털치료기기 개발
기 여 율 90/100
과제수행기관명 주식회사 에스엠디솔루션
연구기간 2021.04.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711126317
과제번호 2020-0-01789-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 High Performance Knowledge System 개발 및 인력양성
기 여 율 10/100
과제수행기관명 동국대학교산학협력단
연구기간 2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

단일 모델의 인공지능 기반으로 복수 언어의 오디오 데이터를 자동 인식하는 다국어 자동음성인식 방법으로서,

음성인식기에 의해, 입력되는 오디오 데이터를 인식하는 단계;

음성언어분류기에 의해, 상기 오디오 데이터를 분류하는 단계;

상기 음성인식기에 결합된 출력계층선택기에 의해, 상기 음성언어분류기로부터 받은 언어 분류 정보에 따라 상기 음성인식기에 각각 연결된 복수의 프로젝션 출력 계층들 중 어느 하나의 프로젝션 출력 계층을 활성화하는 단계-여기서 활성화된 프로젝션 출력 계층의 출력 단위는 1바이트로 설정됨-; 및

상기 활성화된 프로젝션 출력 계층에 의해, 상기 1바이트의 출력 단위로 출력되는 출력들을 재조합하여 상기 오디오 데이터에 대한 자동음성인식 결과로서 출력하는 단계를 포함하는 다국어 자동음성인식 방법.

청구항 2

청구항 1에 있어서,

상기 복수의 프로젝션 출력 계층들은 제1 언어 전용의 제1 프로젝션 출력 계층, 제2 언어 전용의 제2 프로젝션 출력 계층, 및 범용 언어를 위한 제3 프로젝션 출력 계층을 포함하는, 다국어 자동음성인식 방법.

청구항 3

청구항 2에 있어서,

상기 어느 하나의 프로젝션 출력 계층을 활성화하는 단계는, 상기 언어 분류 정보에 포함된 음성 언어 분류의 확실성이 70% 미만인지를 판단하고, 판단 결과, 상기 확실성이 70% 미만일 때, 상기 제3 프로젝션 출력 계층을 활성화하는, 다국어 자동음성인식 방법.

청구항 4

청구항 2에 있어서,

상기 1바이트의 유니코드는 각 언어를 256(2^8)개의 1바이트(8비트)의 조합으로 8진수 표현 방식으로 표현하도록 지원하며,

다국어 자동음성인식 장치에 의해 지원되는 언어의 개수가 증가할 때, 상기 다국어 자동음성인식 장치의 단일 모델의 크기는 256 단계의 출력을 지원하는 프로젝션 출력 계층만 증가하도록 구성되는, 다국어 자동음성인식 방법.

청구항 5

청구항 2에 있어서,

상기 제1 언어는 한국어를 포함하고, 상기 제1 프로젝션 출력 계층은 상기 한국어를 위한 한글의 문자 조합인 2904 가지를 유니코드의 조합을 통해 256개의 바이트로 표현하도록 구성되는, 다국어 자동음성인식 방법.

청구항 6

청구항 1에 있어서,

상기 오디오 데이터를 분류하는 단계는,

음성정보추출기에 의해, 상기 오디오 데이터를 일정 크기 단위로 나누어 음성 정보를 추출하는 단계;

상기 음성정보추출기에 연결된 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들의 평균값

(mean pooling)을 구하는 단계; 및

상기 프로젝터 계층에 연결된 분류기에 의해, 상기 일정 크기 단위로 추출된 음성 정보들의 평균값을 미리 설정된 전용 언어들 중 어느 것에 대응하는지 분류하는 단계를 포함하는, 다국어 자동음성인식 방법.

청구항 7

청구항 6에 있어서,

상기 음성 정보를 추출하는 단계는, 상기 오디오 데이터를 오디오 입력으로 받는 7층의 CNN(convolutional neural network) 구조의 특징 추출부(feature extractor)에 의해 상기 오디오 데이터의 특징을 추출하는 단계를 포함하는, 다국어 자동음성인식 방법.

청구항 8

청구항 7에 있어서,

상기 음성 정보를 추출하는 단계는, 상기 특징 추출부에 연결된 24층의 트랜스포머 인코더(transformer encoder)에 의해, 상기 추출된 특징으로부터 음성 특징 정보 또는 상기 음성 특징 정보에 대응하는 상기 언어 분류 정보를 추출하는 단계를 더 포함하는, 다국어 자동음성인식 방법.

청구항 9

청구항 6에 있어서,

상기 평균값(mean pooling)을 구하는 단계는, 상기 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들에 대해 평균 풀링(mean pooling)을 수행하는, 다국어 자동음성인식 방법.

청구항 10

청구항 6에 있어서,

상기 일정 크기 단위는 25밀리초(ms) 단위인, 다국어 자동음성인식 방법.

청구항 11

단일 모델의 인공지능 기반으로 복수 언어의 오디오 데이터를 자동 인식하는 다국어 자동음성인식 장치로서,

프로세서; 및

상기 프로세서에 의해 실행되는 적어도 하나의 명령을 저장하는 메모리를 포함하되,

상기 프로세서가 실행될 때, 상기 적어도 하나의 명령은 상기 프로세서가:

음성인식기에 의해, 입력되는 오디오 데이터를 인식하는 단계;

음성언어분류기에 의해, 상기 오디오 데이터를 분류하는 단계;

상기 음성인식기에 결합된 출력계층선택기에 의해, 상기 음성언어분류기로부터 받은 언어 분류 정보에 따라 상기 음성인식기에 각각 연결된 복수의 프로젝션 출력 계층들 중 어느 하나의 프로젝션 출력 계층을 활성화하는 단계-여기서 활성화된 프로젝션 출력 계층의 출력 단위는 1바이트로 설정됨-; 및

상기 활성화된 프로젝션 출력 계층에 의해, 상기 1바이트의 출력 단위로 출력되는 출력들을 재조합하여 상기 오디오 데이터에 대한 자동음성인식 결과로서 출력하는 단계를 수행하도록 하는 다국어 자동음성인식 장치.

청구항 12

청구항 11에 있어서,

상기 복수의 프로젝션 출력 계층들은 제1 언어 전용의 제1 프로젝션 출력 계층, 제2 언어 전용의 제2 프로젝션 출력 계층, 및 범용 언어를 위한 제3 프로젝션 출력 계층을 포함하는, 다국어 자동음성인식 장치.

청구항 13

청구항 12에 있어서,

상기 적어도 하나의 명령은 상기 프로세서가, 상기 어느 하나의 프로젝션 출력 계층을 활성화하는 단계에서, 상기 언어 분류 정보에 포함된 음성 언어 분류의 확실성이 70% 미만을 가리킬 때, 상기 제3 프로젝션 출력 계층을 활성화하도록 하는, 다국어 자동음성인식 장치.

청구항 14

청구항 12에 있어서,

상기 1바이트의 유니코드는 각 언어를 256(2^8)개의 1바이트(8비트)의 조합으로 8진수 표현 방식으로 표현하도록 지원하며,

다국어 자동음성인식 장치에 의해 지원되는 언어의 개수가 증가할 때, 상기 다국어 자동음성인식 장치의 음성인식 모델의 크기는 256 단계의 출력을 지원하는 프로젝션 출력 계층만 증가하도록 구성되는, 다국어 자동음성인식 장치.

청구항 15

청구항 12에 있어서,

상기 제1 언어는 한국어를 포함하고, 상기 제1 프로젝션 출력 계층은 상기 한국어를 위한 한글의 문자 조합인 2904 가지를 유니코드의 조합을 통해 256개의 바이트로 표현하도록 구성되는, 다국어 자동음성인식 장치.

청구항 16

청구항 11에 있어서,

상기 적어도 하나의 명령은 상기 프로세서가: 상기 오디오 데이터를 분류하는 단계에서,

음성정보추출기에 의해, 상기 오디오 데이터를 일정 크기 단위로 나누어 음성 정보를 추출하는 단계;

상기 음성정보추출기에 연결된 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들의 평균값(mean pooling)을 구하는 단계; 및

상기 프로젝터 계층에 연결된 분류기에 의해, 상기 일정 크기 단위로 추출된 음성 정보들의 평균값을 미리 설정된 전용 언어들 중 어느 것에 대응하는지 분류하는 단계를 수행하도록 하는, 다국어 자동음성인식 장치.

청구항 17

청구항 16에 있어서,

상기 적어도 하나의 명령은 상기 프로세서가: 상기 음성 정보를 추출하는 단계에서, 상기 오디오 데이터를 오디오 입력으로 받는 7층의 CNN(convolutional neural network) 구조의 특징 추출부(feature extractor)에 의해, 상기 오디오 데이터의 특징을 추출하는 단계를 수행하도록 하는, 다국어 자동음성인식 장치.

청구항 18

청구항 17에 있어서,

상기 적어도 하나의 명령은 상기 프로세서가: 상기 음성 정보를 추출하는 단계에서, 상기 특징 추출부에 연결된 24층의 트랜스포머 인코더(transformer encoder)에 의해, 상기 추출된 특징으로부터 음성 특징 정보 또는 상기 음성 특징 정보에 대응하는 상기 언어 분류 정보를 추출하는 단계를 더 수행하도록 하는, 다국어 자동음성인식 장치.

청구항 19

청구항 16에 있어서,

상기 적어도 하나의 명령은 상기 프로세서가: 상기 평균값(mean pooling)을 구하는 단계에서, 상기 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들에 대해 평균 풀링(mean pooling)을 수행하도록 하는, 다국어 자동음성인식 장치.

청구항 20

청구항 16에 있어서,

상기 일정 크기 단위는 25밀리초(ms) 단위인, 다국어 자동음성인식 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 음성인식 기술에 관한 것으로, 보다 상세하게는 인공지능 모델을 기반으로 다국어의 오디오 데이터를 자동으로 인식하는 음성인식 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 외국인과 내국인이 같이 사용하는 공공기관, 관광지, 안내데스크, 인공지능 비서, 키오스크 등은 다국어 음성인식이 필요한 분야이며, 현재 외국인을 위한 음성인식 시스템이 사용되고 있다. 예를 들어, 여러 국적의 관광객이 모이는 관광지나 공항 등에서 음성인식 기술을 사용하려면, 기본적으로 다국어 음성인식기술이 필요하다.

[0003] 또한, 외국어 학습 소프트웨어에 활용하려면, 외국어를 학습하는 과정에서 직접 발음을 통해 음성인식기가 학습자의 발화를 인식할 수 있도록 하여 학습 중인 단어만 문장을 정확히 발음했는지 판단하는데 사용할 수 있다.

[0004] 인공지능 분야에서 딥러닝 기반의 음성인식 모델은 기본적으로 하나의 언어를 인식가능하게 설계되고 학습되며, 또한 대다수의 다국어를 지원하는 음성인식 시스템은 언어별로 별도의 모델을 사용하여 음성인식을 수행한다. 이러한 기술은 다국어 음성인식기, 인공지능 비서, 인공지능 스피커, 무인 주문 키오스크 등의 사용될 수 있다.

[0005] 종래의 음성인식 시스템은 다국어를 지원하지 않거나, 다국어를 지원하더라도 언어별로 여러 개의 모델을 별도로 사용하게 된다. 또한 언어별 음성인식 모델을 별도로 사용하는 경우, 지원하는 언어의 개수가 많아질 수록 메모리 사용량의 문제가 발생할 수 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위해 도출된 것으로, 본 발명의 목적은, 단일의 인공지능 모델로 한국어를 포함한 다국어의 오디오 데이터를 자동으로 음성인식할 수 있는, 인공지능모델 기반 다국어 음성인식 방법 및 장치를 제공하는데 있다.

[0007] 본 발명의 다른 목적은, 단일 모델로 다국어의 오디오 데이터를 자동 인식하기 위해 언어별 출력 계층을 사용하는, 다국어 음성인식 방법 및 장치를 제공하는데 있다.

[0008] 본 발명의 또 다른 목적은, 음성언어분류기 및/또는 출력계층선택기를 사용하여 언어별 출력 계층들 중 앞서 인식된 음성언어에 대응하는 특정 출력 계층만을 활성화하여 단일 모델로 다국어의 오디오 데이터를 자동 인식할 수 있는, 다국어 음성인식 방법 및 장치를 제공하는데 있다.

[0009] 본 발명의 또 다른 목적은, 저사양의 키오스크, 퍼스널 컴퓨터(personal computer, PC), 컴퓨터 단말기 등에서도 1개 모델만큼의 메모리만으로 다국어 음성인식을 수행할 수 있는, 다국어 음성인식 방법 및 장치를 제공하는데 있다.

과제의 해결 수단

[0010] 전술한 목적을 달성하기 위한 본 발명의 일 측면에 따른 다국어 자동음성인식 방법은, 단일 모델의 인공지능 기반으로 복수 언어의 오디오 데이터를 자동 인식하는 다국어 자동음성인식 방법으로서, 음성인식기에 의해, 입력되는 오디오 데이터를 인식하는 단계; 음성언어분류기에 의해, 상기 오디오 데이터를 분류하는 단계; 상기 음성인식기에 결합된 출력계층선택기에 의해, 상기 음성언어분류기로부터 받은 언어 분류 정보에 따라 상기 음성인식기에 각각 연결된 복수의 프로젝션 출력 계층들 중 어느 하나의 프로젝션 출력 계층을 활성화하는 단계-여기서 활성화된 프로젝션 출력 계층의 출력 단위는 1바이트로 설정됨-; 및 상기 활성화된 프로젝션 출력 계층에 의해, 상기 1바이트의 출력 단위로 출력되는 출력들을 재조합하여 상기 오디오 데이터에 대한 자동음성인식 결과

로서 출력하는 단계를 포함한다.

[0011] 일실시예에서, 상기 어느 하나의 프로젝션 출력 계층을 활성화하는 단계는, 상기 언어 분류 정보에 포함된 음성 언어 분류의 확실성이 70% 미만인지를 판단하고, 상기 확실성이 70% 미만일 때, 상기 제3 프로젝션 출력 계층을 활성화할 수 있다.

[0012] 일실시예에서, 상기 오디오 데이터를 분류하는 단계는, 음성정보추출기에 의해, 상기 오디오 데이터를 일정 크기 단위로 나누어 음성 정보를 추출하는 단계; 상기 음성정보추출기에 연결된 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들의 평균값(mean pooling)을 구하는 단계; 및 상기 프로젝터 계층에 연결된 분류기에 의해, 상기 일정 크기 단위로 추출된 음성 정보들의 평균값을 미리 설정된 전용 언어들 중 어느 것에 대응하는지 분류하는 단계를 포함할 수 있다.

[0013] 일실시예에서, 상기 음성 정보를 추출하는 단계는, 상기 오디오 데이터를 오디오 입력으로 받는 7층의 CNN(convolutional neural network) 구조의 특징 추출부(feature extractor)에 의해 상기 오디오 데이터의 특징을 추출하는 단계를 포함할 수 있다.

[0014] 일실시예에서, 상기 음성 정보를 추출하는 단계는, 상기 특징 추출부에 연결된 24층의 트랜스포머 인코더(transformer encoder)에 의해, 상기 추출된 특징으로부터 음성 특징 정보 또는 상기 음성 특징 정보에 대응하는 상기 언어 분류 정보를 추출하는 단계를 더 포함할 수 있다.

[0015] 일실시예에서, 상기 평균값(mean pooling)을 구하는 단계는, 상기 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들에 대해 평균 풀링(mean pooling)을 수행하도록 구성될 수 있다.

[0016] 전술한 목적을 달성하기 위한 본 발명의 다른 측면에 따른 다국어 자동음성인식 장치는, 단일 모델의 인공지능 기반으로 복수 언어의 오디오 데이터를 자동 인식하는 다국어 자동음성인식 장치로서, 프로세서; 및 상기 프로세서에 의해 실행되는 적어도 하나의 명령을 저장하는 메모리를 포함하되, 상기 프로세서가 실행될 때, 상기 적어도 하나의 명령은 상기 프로세서가: 음성인식기에 의해, 입력되는 오디오 데이터를 인식하는 단계; 음성언어 분류기에 의해, 상기 오디오 데이터를 분류하는 단계; 상기 음성인식기에 결합된 출력계층선택기에 의해, 상기 음성언어분류기로부터 받은 언어 분류 정보에 따라 상기 음성인식기에 각각 연결된 복수의 프로젝션 출력 계층들 중 어느 하나의 프로젝션 출력 계층을 활성화하는 단계-여기서 활성화된 프로젝션 출력 계층의 출력 단위는 1바이트로 설정됨-; 및 상기 활성화된 프로젝션 출력 계층에 의해, 상기 1바이트의 출력 단위로 출력되는 출력들을 재조합하여 상기 오디오 데이터에 대한 자동음성인식 결과로서 출력하는 단계를 수행하도록 구성될 수 있다.

[0017] 일실시예에서, 상기 적어도 하나의 명령은 상기 프로세서가, 상기 어느 하나의 프로젝션 출력 계층을 활성화하는 단계에서, 상기 언어 분류 정보에 포함된 음성 언어 분류의 확실성이 70% 미만을 가리킬 때, 상기 제3 프로젝션 출력 계층을 활성화하도록 구성될 수 있다.

[0018] 일실시예에서, 상기 적어도 하나의 명령은 상기 프로세서가: 상기 음성 정보를 추출하는 단계에서, 상기 특징 추출부에 연결된 24층의 트랜스포머 인코더(transformer encoder)에 의해, 상기 추출된 특징으로부터 음성 특징 정보 또는 상기 음성 특징 정보에 대응하는 상기 언어 분류 정보를 추출하는 단계를 더 수행하도록 구성될 수 있다.

[0019] 일실시예에서, 상기 적어도 하나의 명령은 상기 프로세서가: 상기 평균값(mean pooling)을 구하는 단계에서, 상기 프로젝터 계층에 의해, 상기 일정 크기 단위로 추출된 음성 정보들에 대해 평균 풀링(mean pooling)을 수행하도록 구성될 수 있다.

[0020] 일실시예에서, 상기 복수의 프로젝션 출력 계층들은 제1 언어 전용의 제1 프로젝션 출력 계층, 제2 언어 전용의 제2 프로젝션 출력 계층, 및 범용 언어를 위한 제3 프로젝션 출력 계층을 포함한다.

[0021] 일실시예에서, 상기 1바이트의 유니코드는 각 언어를 256(2^8)개의 1바이트(8비트)의 조합으로 8진수 표현 방식으로 표현하도록 사용될 수 있다.

[0022] 일실시예에서, 다국어 자동음성인식 장치에 의해 지원되는 언어가 1개 증가할 때, 다국어 자동음성인식 장치의 단일 모델의 크기는 256 단계의 출력을 지원하는 프로젝션 출력 계층만 1개 증가하도록 구성될 수 있다.

[0023] 일실시예에서, 상기 제1 언어는 한국어를 포함하고, 상기 제1 프로젝션 출력 계층은 상기 한국어를 위한 한글의 문자 조합인 2904 가지를, 유니코드의 조합을 통해 256개의 바이트로 표현하도록 구성될 수 있다.

[0024] 일실시예에서, 상기 일정 크기 단위는 25밀리초(ms) 단위일 수 있다.

발명의 효과

[0025] 본 발명에 의하면, 종래 기술의 단일 언어만 지원하는 음성인식 모델과 다르게, 즉 복수의 언어에는 복수의 모델들을 필요로 하는 종래 기술과 달리, 단일 모델로 다국어 자동음성인식이 가능한 장점이 있다. 특히, 음성인식 모델에 언어별로 별도의 출력계층을 추가하여 사용하도록 구성함으로써, 단일 모델이 다수의 모델을 이용한 음성인식 시스템에 비해 인식률이 떨어지는 문제를 해결하여 효율적인 다국어 자동음성인식 시스템을 구현할 수 있다.

[0026] 또한, 본 발명에 의하면, 다국어를 지원할 때, 사용자가 음성 인식을 시작하기 전에 인식할 언어를 선택해야 하는 과정을 생략할 수 있다. 또한, 언어를 분류하는 과정을 추가해 사용자가 명시적으로 언어를 선택하지 않아도 어떤 언어인지를 자동으로 구분해낼 수 있다.

[0027] 또한, 본 발명에 의하면, 언어 분류 정보를 이용해 음성인식 모델 내 특정 언어 전용 출력 계층을 활성화하여 해당 언어를 인식하도록 함으로써 성능을 증진시킬 수 있다. 또한, 음성 언어의 분류의 확실성이 낮은 경우에는 범용 출력 계층을 활용하하여 사용함으로써 자동 음성 인식에 대한 오작동을 방지할 수 있다.

[0028] 또한, 본 발명에 의하면, 기존 단일어 음성인식기와 다르게 다국어를 하나의 모델에서 동시에 인식할 수 있기 때문에, 다양한 장치에서 1개 모델만큼의 메모리만으로 다국어 음성인식을 수행할 수 있다. 특히, 저사양의 키오스크, 퍼스널 컴퓨터(personal computer, PC), 컴퓨터 단말기 등에서 하나의 딥러닝 모델만으로 다국어 음성인식이 가능한 효과가 있다. 아울러, 음향 모델을 공유하여 저자원의 각각의 언어의 음성인식 성능을 개선시킬 수 있다.

도면의 간단한 설명

[0029] 도 1은 본 발명의 일 실시예에 따른 다국어 자동음성인식 장치의 전체 구조에 대한 개략적인 블록도이다.

도 2는 도 1의 다국어 자동음성인식 장치의 음성인식기에 채용할 수 있는 음성인식 모델을 설명하기 위한 블록도이다.

도 3은 도 1의 다국어 자동음성인식 장치의 음성언어분류기에 채용할 수 있는 인공지능 모델을 설명하기 위한 블록도이다.

도 4는 도 3의 음성언어분류기의 프로젝터 계층의 작동 원리를 설명하기 위한 블록도이다.

도 5는 본 발명의 다른 실시예에 따른 다국어 자동음성인식 장치에 대한 개략적인 블록도이다.

도 6은 도 5의 다국어 자동음성인식 장치에 채용할 수 있는 특정 작동 원리를 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0030] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하여 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0031] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. '및/또는'이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0032] 본 출원의 실시예들에서, 'A 및 B 중에서 적어도 하나'는 'A 또는 B 중에서 적어도 하나' 또는 'A 및 B 중 하나 이상의 조합들 중에서 적어도 하나'를 의미할 수 있다. 또한, 본 출원의 실시예들에서, 'A 및 B 중에서 하나 이상'은 'A 또는 B 중에서 하나 이상' 또는 'A 및 B 중 하나 이상의 조합들 중에서 하나 이상'을 의미할 수 있다.

[0033] 어떤 구성요소가 다른 구성요소에 '연결되어' 있다거나 '접속되어' 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이

해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 '직접 연결되어' 있다거나 '직접 접속되어'있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0034] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, '포함한다' 또는 '가진다' 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0035] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0036] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0037] 도 1은 본 발명의 일 실시예에 따른 다국어 자동음성인식 장치의 전체 구조에 대한 개략적인 블록도이다.

[0038] 도 1을 참조하면, 다국어 자동음성인식 장치(이하 간략히 '음성인식 장치)(100)는 단일 인공지능 모델을 구비할 수 있고, 좀더 구체적으로는 음성인식기(speech recognitor, 10) 및 언어별 출력 계층(40)을 구비할 수 있다.

[0039] 음성인식기(10)는, 간략히 ASR(10)으로 지칭될 수 있고, 7층 CNN(Convolutional Neural Network) 구조의 특징 추출부(feature extractor)에 의해 오디오 데이터로부터 특징을 추출한 후, 추출된 특징으로부터 24층 트랜스포머 인코더(transformer encoder)에 의해 음성 특징 정보를 추출하도록 구성될 수 있다. 이러한 음성인식기(10)는 Wav2Byte 음성인식 모델로 지칭될 수 있다.

[0040] 언어별 출력 계층(40)은 제1 언어 전용의 제1 프로젝션 계층(41), 제2 언어 전용의 제2 프로젝션 계층(42) 및 범용 프로젝션 계층(44)을 구비할 수 있다. 각 프로젝션 계층(projection layer)은 프로젝션 출력 계층으로 각각 지칭될 수 있다.

[0041] 또한, 언어별 출력 계층(40)은 중국어, 스페인어, 영어, 힌디어, 아랍어, 벵골어, 포르투갈어, 러시아어, 일본어, 란다어, 마라티어, 텔루구어, 말레이어, 터키어, 한국어, 프랑스어, 독일어, 베트남어, 타밀어, 우르두어, 자바어, 이탈리아어, 페르시아어, 구자라트어, 보즈푸리어 등에서 선택되는 어느 하나의 특정 언어를 위한 제3 프로젝션 계층(43)을 더 구비할 수 있다.

[0042] 본 실시예에서 제1 언어 내지 제3 언어는 서로 다른 언어들이며, 예를 들어, 제1 언어는 한국어, 제2 언어는 영어, 그리고 제3 언어는 일본어일 수 있다.

[0043] 전술한 음성인식 장치(100)는, 단일 모델의 인공지능 기반으로 복수 언어의 오디오 데이터를 자동 음성인식(automatic speech recognition)하도록 구성될 수 있다. 이를 위해, 음성인식 장치(100)는, 입력되는 오디오(audio) 데이터를 인식하고, 인식에 따라 활성화된 특정 프로젝션 출력 계층에 의해 오디오 데이터로부터 1바이트의 출력 단위로 생성되는 출력들을 재조합하여 자동음성인식 결과(ARS result)로 출력할 수 있다.

[0044] 즉, 음성인식 장치(100)는, 프로젝션 출력 계층의 출력 단위를 1바이트로 설정함으로써, 1바이트 출력 단위의 256개 출력들을 재조합하여 자동음성인식 결과를 생성하도록 구성될 수 있다.

[0045] 전술한 음성인식기(10)와 언어별 출력 계층(40)의 구성에 의하면, 단일 모델 예컨대 단일 인공지능 모델을 사용하여 다국어의 오디오 데이터에 대한 음성인식을 자동으로 수행할 수 있다.

[0046] 한편, 전술한 음성인식 장치(100)에서는 입력되는 오디오 데이터에 따라 특정 프로젝션 출력 계층을 활성하기 위해 음성언어를 분류하기 구성부를 별도록 구성할 수 있다. 즉, 음성인식 장치(100)는 음성언어분류기(20)을 더 구비할 수 있다. 음성언어분류기(20)는 간략히 분류기(classifier, 20)으로 지칭될 수 있다.

[0047] 음성언어분류기(20)는 뉴럴 네트워크 기반 피처 추출 기법들인 Wav2Vec 모델 및 VQ-Wav2Vec 모델 중 어느 하나를 기반으로 형성될 수 있다.

[0048] Wav2Vec 모델은 크게 인코더 네트워크(encoder network)와 컨텍스트 네트워크(context network)로 구성된다. 이 둘 모두는 컨볼루션 뉴럴 네트워크(convolutional neural network, CNN)일 수 있다. 여기서, 인코더 네트워크는 음성 입력을 히든 리프리젠테이션(hidden representaion)으로 인코딩하는 역할을 수행하고, 컨텍스트 네트워크는 히든 리프리젠테이션을 컨텍스트 리프레젠테이션(context representation)으로 변환하는 역할을 수행한다. Wav2Vec 모델의 학습이 완료되면, 컨텍스트 리프리젠테이션을 해당 오디오 데이터의 특징(feature)로 사용할 수 있다.

[0049] 즉, Wav2Vec 모델은 오디오 데이터의 입력이 포지티브 쌍인지 네거티브 쌍인지 이진 분류(binary classification)하는 과정에서 학습될 수 있다. 포지티브 쌍은 입력 오디오 데이터의 i번째 컨텍스트 리프리젠테이션과 i+1번째 히든 리프리젠테이션으로 구성되고, 네거티브 쌍은 입력 오디오 데이터의 i번째 컨텍스트 리프리젠테이션과 현재 배치의 다른 오디오 데이터의 히든 리프리젠테이션들 가운데 랜덤으로 추출한 것으로 구성될 수 있다.

[0050] Wav2Vec 모델의 학습이 진행될수록 포지티브 쌍 관계의 리프리젠테이션은 벡터 공간에서 가까워지고, 네거티브 쌍 관계의 리프리젠테이션은 벡터 공간에서 멀어진다. 다시 말해서, 인코더 네트워크와 컨텍스트 네트워크는 입력 오디오 데이터의 다음 시퀀스가 무엇일지에 관한 정보를 오디오 데이터의 특징에 잘 녹여낼 수 있다.

[0051] VQ-Wav2Vec 모델은 그 중간에 벡터 양자화(vector quantization, VQ) 모듈이 추가된 것으로 제외하고 전술한 Wav2Vec 모델과 그 아키텍처가 동일할 수 있다. 벡터 양자화 모듈은 검벨 소프트맥스(Gumbel softmax) 방식이나 K-평균 클러스터링(K-means clustering) 방식을 채용하여 구성될 수 있다. 검벨 소프트맥스 방식의 VQ-Wav2Vec 모델은, 히든 리프리젠테이션을 선형변환하여 로짓(logit)을 만든 후, 로짓으로부터 검벨 소프트맥스와 argmax를 순차적으로 취해 원핫 벡터를 만들고, 연속적인 변수인 히든 리프리젠테이션을 원핫 벡터와 임베딩 매트릭스로 내적하여 히든 리프리젠테이션을 복수의 임베딩 중 선택된 하나의 이산(discrete) 변수로 변환한 것을 지칭할 수 있다.

[0052] 이와 같이, 음성언어분류기(20)는 CNN과 트랜스포머들(transformers)을 결합하여 구성할 수 있다.

[0053] 또한, 전술한 음성언어분류기(20)는 음성인식기(10)의 음성 입력시에 전달되는 동기화신호 또는 제어신호(S11)에 의해 트리거되어 입력 오디오 데이터가 어느 언어에 속하는 음성인지를 분류하고, 분류된 음성에 대한 언어 분류 정보를 포함한 피드 언어 정보(feed language information, FLI)를 출력하도록 구성될 수 있다.

[0054] 한편, 전술한 구성에 있어서, 다국어 자동음성인식 장치(100)는 피드 언어 정보(FLI)를 토대로 복수의 출력 계층들 중 특정 출력 계층을 활성화하기 위한 출력계층선택기(30)를 더 구비할 수 있다.

[0055] 출력계층선택기(30)는 피드 언어 정보의 신호 레벨이나 저장값에 따라 하나의 출력 계층을 활성화하도록 구성될 수 있다. 이러한 출력계층선택기(30)는 언어별 출력 계층들(40) 중 하나의 출력 계층을 활성화하기 위해, 음성인식기(10)의 출력단이나, 언어별 출력 계층들(40)의 입력단이나, 이들 사이에 설치될 수 있다.

[0056] 즉, 출력계층선택기(30)는 음성인식기(10)의 하나의 특정 출력단을 활성화하거나 연결하도록 구성될 수 있고, 이와 유사하게 언어별 출력 계층들(40) 중 하나의 특정 출력 계층의 입력단을 활성화하거나 연결하도록 구성될 수 있다. 또한, 대안적으로, 출력계층선택기(30)는 음성인식기(10)의 출력단과 언어별 출력 계층들(40)의 각 입력단과의 사이를 연결하기 위해 설치된 복수의 배선들이나 논리적인 채널들 중 어느 하나만을 활성화시키도록 구성될 수 있다.

[0057] 전술한 구성에 의하면, 인공지능 모델은, 입력으로 받은 오디오 데이터를 음성인식기(10)와 음성언어분류기(20)에 각각 입력하고, 음성언어분류기(20)의 출력인 언어 분류 정보(FLI)를 이용해 음성인식기(10)의 언어별 CTC(connectionist temporal classification) 프로젝션 출력 계층들 중 특정 출력 계층을 선택하여 최종 음성인식 결과(ARS result)를 생성할 수 있다.

[0058] 또한, 전술한 구성에 있어서, 음성 언어 분류의 확실성(confidence)이 70% 미만인 경우에는 음성인식 장치(100)의 오작동을 방지하기 위하여 전용 언어 출력 계층이 아닌 범용(ALL) 언어 출력 계층을 이용해 음성 인식 결과를 생성할 수 있다.

[0059] 이와 같이 본 실시예에서는 입력 오디오 데이터의 음성언어를 분류하는 분류기(classifier, 20)를 이용해 음성의 언어를 분류하고, 분류된 언어를 이용해 음성인식기(ASR, 10)가 출력할 언어를 선택할 수 있다. 또한, 음성 분류의 확실성이 낮은 경우 음성인식기(10)가 지원하는 모든 언어를 출력하는 범용 출력계층(44)을 이용해 음성인식 결과를 출력할 수 있다.

[0060] 즉, 음성인식기의 구조는 다국어 음성인식을 위해 제안된 인공지능 모델 구조로서 인식하는 모든 언어가 Wav2Byte 모델을 공통적으로 공유하도록 구성되고, 각 언어별로 별도의 CTC 프로젝션 출력 계층을 가지도록 구성된다. 또한, 음성언어분류기를 통해 사용자의 발화가 어느 언어인지 판별해 음성인식 단계에서 해당 언어의 CTC 프로젝션 출력 계층을 활성화하여 최적의 음성인식 결과를 출력할 수 있다. 특히, CTC 프로젝션 출력 계층의 출력 단위를 1바이트(8비트)에 해당하는 유니코드(unicode) 단위로 설정하고, 이러한 유니코드 복수개로 오디오 데이터의 단위 출력들을 생성하여 재조합함으로써 모든 언어에 대해 적용하여 음성인식 결과를 생성할 수 있다. 더욱이, 음성언어 분류가 확실하지 않은 경우에, 모든 언어를 동시에 학습하고, 모든 언어의 출력이 가능한 범용 출력 계층을 사용하여 음성인식 결과를 출력하도록 구성함으로써, 전용 출력 계층보다는 성능이 조금 떨어지지만, 음성 언어 인식이 제대로 이뤄지지 않을 때 잘못된 언어의 출력을 생성하는 오작동을 방지할 수 있다.

[0061] 또한, 본 실시예의 음성인식 장치는, 다국어 음성인식이 필요한 상황에서 단말기기의 하드웨어적 제약으로 인해 단일어 정도만 음성인식이 가능한 사양을 가진 경우에 유용하게 적용할 수 있다. 예를 들어, 저사양 음성인식 장치에 설치된 하나의 인공지능 모델로 다국어 음성인식을 수행하고, 자동으로 사용자의 발화를 인식하여 처리할 수 있는 음성인식 서비스를 효율적으로 구현할 수 있다.

[0062] 전술한 음성인식기에 채용할 수 있는 음성인식 모델을 좀더 구체적으로 살펴보면 다음과 같다.

[0063] 도 2는 도 1의 다국어 자동음성인식 장치의 음성인식기에 채용할 수 있는 음성인식 모델을 설명하기 위한 블록도이다.

[0064] 도 2를 참조하면, 본 실시예의 음성인식 모델은 한국어와 영어를 인식하는 모델 형태를 가질 수 있다.

[0065] 본 실시예의 음성인식기(10)의 음성인식 모델은, 영어(EN)와 한국어(KO)의 오디오 입력을 7층 CNN(convolutional neural network) 구조의 특징 추출부(feature extractor, 12)와 24층 CNN 구조의 트랜스포머 인코더(transformer encoder, 14)를 통과시켜 음성 특징 정보를 추출하도록 구성될 수 있다.

[0066] 그런 다음, 음성인식 모델은, 한국어(KO)를 출력할 수 있는 CTC 프로젝션(projection) 출력 계층(41), 영어(EN)를 출력할 수 있는 CTC 프로젝션 출력 계층(42), 혹은 한국어와 영어를 모두(all) 출력할 수 있는 CTC 프로젝션 출력 계층(44)을 통해 최종 음성인식 결과를 출력할 수 있가.

[0067] 이때, 각 CTC 프로젝션 출력 계층(21, 42, 44)의 출력 단위는 1바이트(8bits)에 해당하는 유니코드(unicode) 단위이고, 음성인식 모델을 복수개의 출력 단위들을 재조합해 최종 결과를 생성할 수 있다. 복수개는 256개가 바람직하나, 이에 한정되지 않을 수 있다. 여기서, 바이트 단위의 정보는 8진수 표현 방식으로, 예컨대 0x00 내지 0xff 내의 값들로 표현될 수 있다.

[0068] 바이트 단위의 유니코드 표현은 전세계의 모든 언어를 256(2^8)개의 8비트(1바이트) 의 조합으로 표현할 수 있어, 음성인식 모델이 지원하는 언어의 개수가 증가하더라도 전체 모델의 크기는 256 단계의 출력을 지원하는 CTC 프로젝션 출력 계층만큼만 증가하게 된다.

[0069] 특히, 한국어의 경우, 한글이 가질 수 있는 2904 가지의 문자 조합에 대하여 음성인식 모델을 생성하는 데에는 상당히 출력 계층이 필요하게 되지만, 본 실시예의 유니코드의 조합을 통해 표현하면 모두 256개의 바이트로 표현할 수 있다.

[0070] 표 1은 다국어를 2진수 유니코드(byte)로 표현하는 예시이다.

표 1

[0071]

Charcters	Bytes
A	010000001
가	11101010 10110000 10000000
안녕	11101100 10010101 10001000 11101011 10000101 10010101
あ	11100011 10000001 10000010

[0072] 이와 같이, 한국어, 영어, 일본어 외에 전세계의 대부분의 언어를 유니코드의 조합을 통해 256개의 바이트로 각각 표현할 수 있다.

[0073] 본 실시예에 의하면, 음성인식 장치에 인식되는 모든 언어가 공통적으로 음성인식 모델 구조를 공유하도록 하고, 각 언어별로 별도의 CTC 프로젝션 출력 계층을 가지도록 설정하면서, 음성언어 분류가 확실하지 않은 경우에, 모든 언어를 동시에 학습하여 출력할 수 있는 범용 출력 계층을 사용하도록 설정함으로써, 음성인식의 오작동을 방지하여 음성인식 장치의 신뢰성을 크게 증대시킬 수 있다.

[0074] 한편, 도 1을 참조하여 앞서 설명한 음성언어분류기에 채용할 수 있는 음성언어 분류 모델을 좀더 구체적으로 살펴보면 다음과 같다.

[0075] 도 3은 도 1의 다국어 자동음성인식 장치의 음성언어분류기에 채용할 수 있는 음성언어 분류 모델을 설명하기 위한 블록도이고, 도 4는 도 3의 음성언어 분류 모델의 프로젝터 계층의 작동 원리를 설명하기 위한 블록도이다.

[0076] 도 3을 참조하면, 본 실시예의 음성언어분류기(20)의 음성언어 분류 모델은 음성인식 모델의 음성인식 성능을 증진시키기 위해 탑재될 수 있다.

[0077] 음성언어 분류 모델은 음성정보추출기, 프로젝터(projector) 계층(26) 및 분류기(classifier, 28)를 구비할 수 있다. 음성정보추출기는 Wav2Vec 모델을 기반으로 구성될 수 있고, 특징 추출부(feature extractor, 22) 및 트랜스포머 인코더(transformer encoder, 24)를 구비할 수 있다.

[0078] 전술한 음성언어 분류 모델은, 음성 정보를 25밀리초(ms) 단위로 나누어 특징 추출부(22)와 트랜스포머 인코더(24)를 이용해 음성 정보를 추출하고, 프로젝터 계층(26)을 통해 25밀리초 단위의 추출된 음성 정보들의 평균값을 연산하고, 분류기(28)에 의해 음성 정보에 대한 언어를 분류할 수 있다.

[0079] 한편, 프로젝터 계층(26)은, 도 4에 도시한 바와 같이, 25밀리초 단위의 음성 정보들(261, 262, 263, 264)의 평균 풀링(mean pooling) 연산을 통해 하나의 음성 정보(pooled unit, 280)를 생성할 수 있다.

[0080] 도 5는 본 발명의 다른 실시예에 따른 다국어 자동음성인식 장치에 대한 개략적인 블록도이다.

[0081] 도 5를 참조하면, 음성인식 장치(100)는, 프로세서(processor, 110) 및 프로세서(110)에 의해 실행되는 적어도 하나의 명령을 저장하는 메모리(memory, 120)를 포함하여 구성될 수 있다. 또한, 음성인식 장치(100)는 송수신 장치(transceiver, 130), 저장 장치(storage, 140), 입력 인터페이스 장치(150), 출력 인터페이스 장치(160) 및 버스(bus)를 포함하여 구성될 수 있다.

[0082] 적어도 하나의 명령은 프로세서(110)가 다국어 음성인식이 필요한 상황에서 단말기기의 하드웨어적 제약으로 인해 단일어만 음성인식이 가능한 사양을 가진 경우에도, 하나의 인공지능 모델로 다국어 음성인식을 수행하고, 자동으로 사용자의 발화를 인식하여 음성인식 결과를 제공하는 서비스를 수행하도록 구성될 수 있다.

[0083] 프로세서(110)는 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 발명의 실시예들에 따른 방법들이 수행되는 전용의 프로세서를 의미할 수 있다.

[0084] 메모리(120) 및 저장 장치(140) 각각은 휘발성 저장 매체 및 비휘발성 저장 매체 중에서 적어도 하나로 구성될 수 있다. 예를 들어, 메모리(120) 및 저장 장치(140) 각각은 읽기 전용 메모리(read only memory, ROM) 및 랜덤 액세스 메모리(random access memory, RAM) 중에서 적어도 하나로 구성될 수 있다.

[0085] 송수신 장치(130)는 유선, 무선 또는 위성 네트워크를 통해 통신을 수행하기 위한 적어도 하나의 서브통신시스템을 구비할 수 있다.

[0086] 전술한 본 실시예의 음성인식 장치(100)는, 예를 들어, 데스크탑 컴퓨터(desktop computer), 랩탑 컴퓨터(laptop computer), 노트북(notebook), 스마트폰(smart phone), 태블릿 PC(tablet PC), 모바일폰(mobile phone), 스마트 워치(smart watch), 스마트 글래스(smart glass), e-book 리더기, PMP(portable multimedia player), 휴대용 게임기, 네비게이션(navigation) 장치, 디지털 카메라(digital camera), DMB(digital multimedia broadcasting) 재생기, 디지털 음성 녹음기(digital audio recorder), 디지털 음성 재생기(digital audio player), 디지털 동영상 녹화기(digital video recorder), 디지털 동영상 재생기(digital video player), PDA(Personal Digital Assistant) 등에 일체로 결합되거나 탑재될 수 있다.

[0087] 한편, 본 실시예의 변형예에서는 도 1에 도시한 음성인식 장치의 구조에 더하여 음성언어분류기의 분류 결과가 한국어인 경우, 음성인식기의 입력단 앞에서 입력 오디오 데이터의 자소를 분리하는 전처리를 수행하기 위한 전처리 수단이나 이러한 전처리 수단에 상응하는 기능을 수행하는 구성부를 더 포함하도록 구성될 수 있다.

[0088] 도 6은 도 5의 음성인식 장치에 채용할 수 있는 특정 작동 원리를 설명하기 위한 흐름도이다.

[0089] 도 6을 참조하면, 음성인식 장치는, 음성인식기에 음성입력이 인식될 때 제1 제어신호를 음성언어 분류기로 전달할 수 있다(S61).

[0090] 다음, 음성언어 분류기의 언어 분류 정보를 출력 계층 선택기로 전달할 수 있다(S63).

[0091] 다음, 음성언어 분류기의 음성언어 분류에 대한 확실성이 70% 미만인지를 판단할 수 있다.

[0092] 확실성이 70% 이상이면, 음성인식 장치는 음성언어 분류기의 음성 분류 정보를 토대로 출력 계층 선택기의 동작에 따라 활성화되는 출력 계층을 통해 해당 언어의 음성인식 결과를 출력할 수 있다(S67).

[0093] 한편, 확실성이 70% 미만이면, 음성인식 장치는 음성언어 분류기의 음성 분류 정보를 무시하고, 음성인식 장치에서 지원하는 모든 언어를 출력할 수 있는 범용 언어 출력 계층 즉, 범용 출력 계층을 통해 음성입력에 대한 음성인식 결과를 출력할 수 있다(S69).

[0094] 전술한 음성인식 장치에 의하면, 기존 단일어 음성인식기와 다르게 하나의 인공지능 모델에서 다국어를 동시에 인식할 수 있기 때문에, 다양한 장치에서 모델 1개만큼의 메모리만으로 다국어 음성인식을 용이하게 구현할 수 있다. 즉, 저사양의 키오스크, 퍼스널 컴퓨터, 각종 전자기기 또는 단말기 등에서 하나의 딥러닝 모델만으로 다국어 음성인식을 구현할 수 있는 효과가 있다. 또한, 공유된 음성인식 모델을 통해 저자원에서 다중 언어의 음성인식 성능을 개선할 수 있다.

[0095] 본 발명의 실시예에 따른 음성인식 방법의 동작은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

[0096] 또한, 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

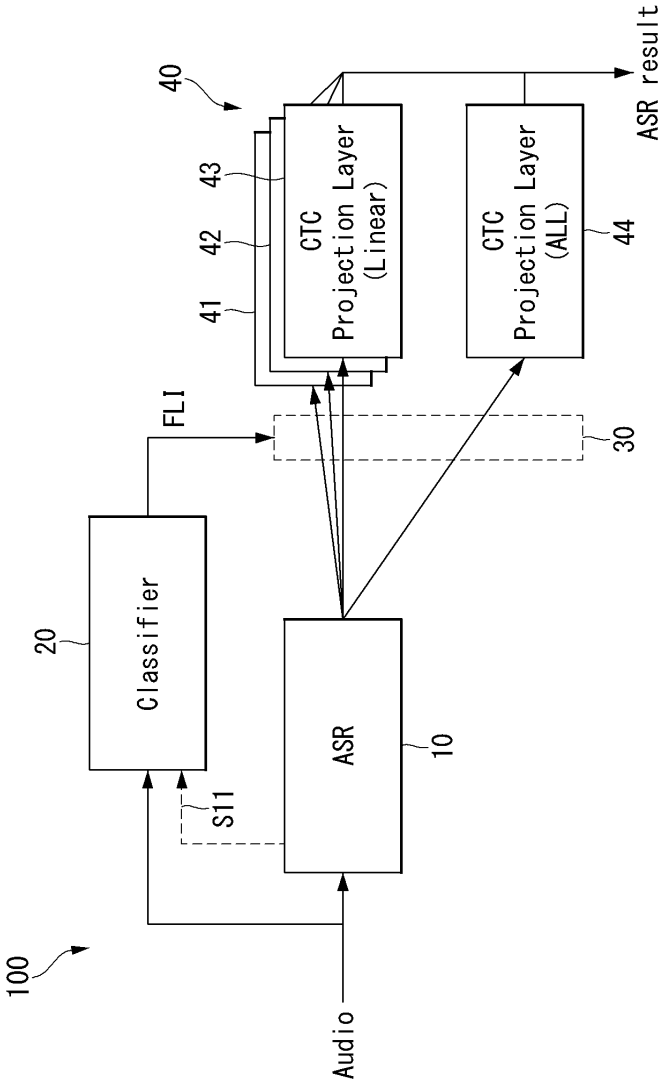
[0097] 본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시예에서, 가장 중요한 방법 단계들의 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0098] 실시예들에서, 프로그램 가능한 로직 장치(예를 들어, 필드 프로그래머블 게이트 어레이)가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시예들에서, 필드 프로그래머블 게이트 어레이는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

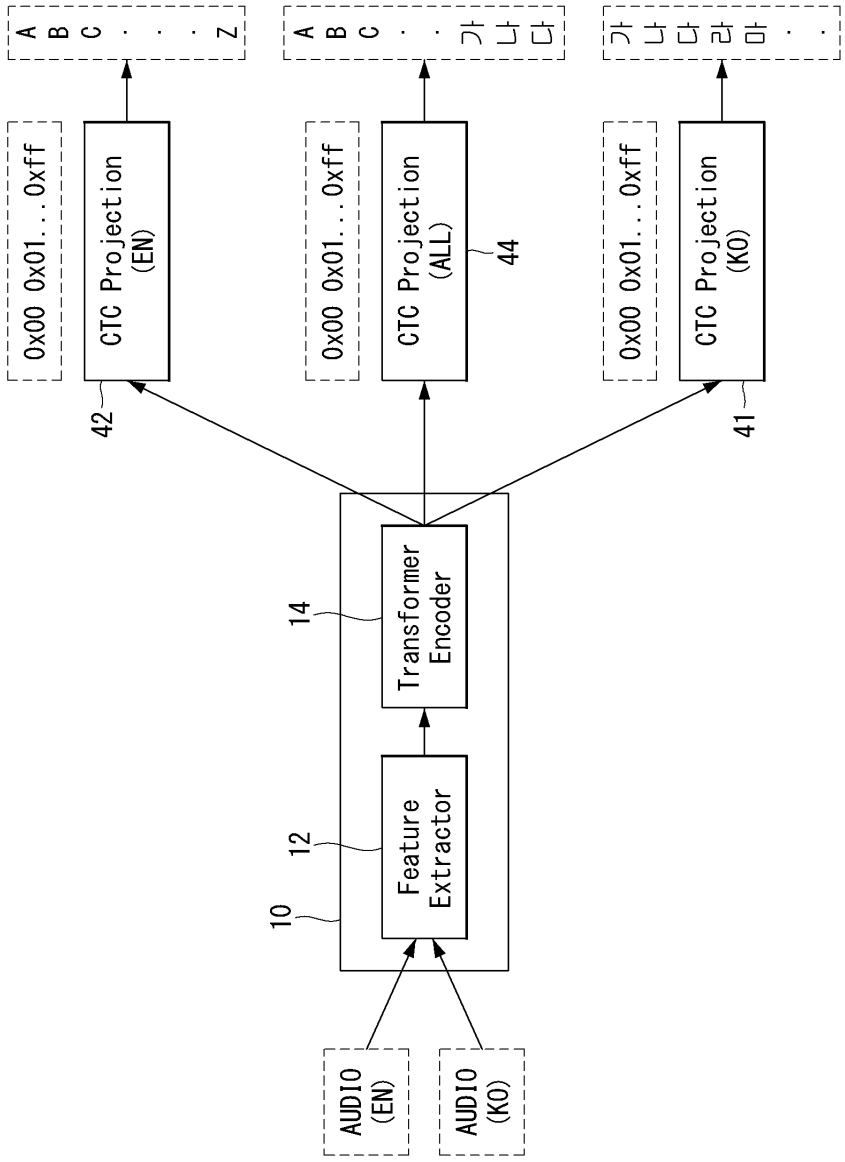
[0099] 이상 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 청구범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

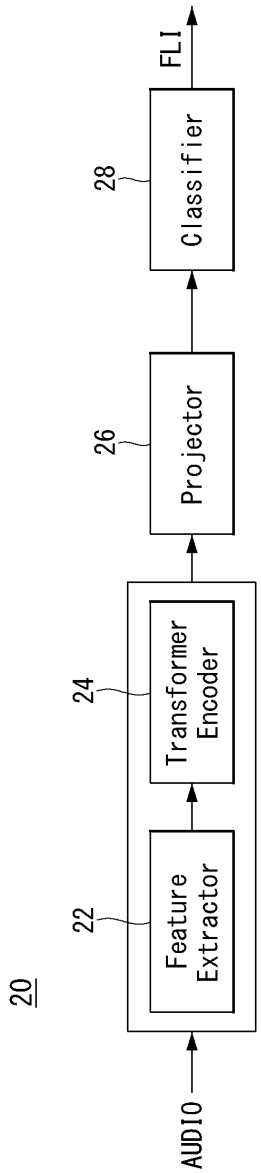
도면1


100
Audio
ASR
10
Classifier
20
S11
FLI
30
40
41
42
43
CTC
Projection Layer
(Linear)
CTC
Projection Layer
(ALL)
44
ASR result

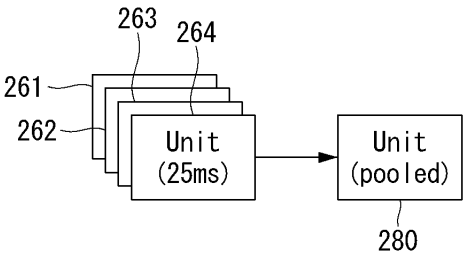
도면2


A B C · · · Z
A B C · · 가 나 다
가 나 다 라 마 · ·
0x00 0x01...0xff
0x00 0x01...0xff
0x00 0x01...0xff
CTC Projection (EN)
CTC Projection (ALL)
CTC Projection (KO)
42
44
41
14
Transformer Encoder
12
Feature Extractor
10
AUDIO (EN)
AUDIO (KO)

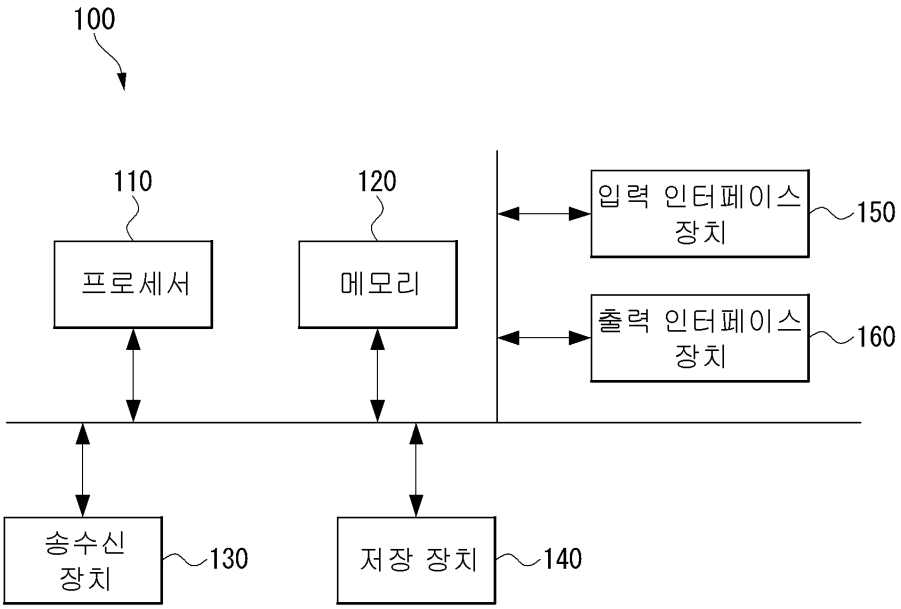
도면3


FLI
28
Classifier
26
Projector
24
Transformer Encoder
22
Feature Extractor
20
AUDIO

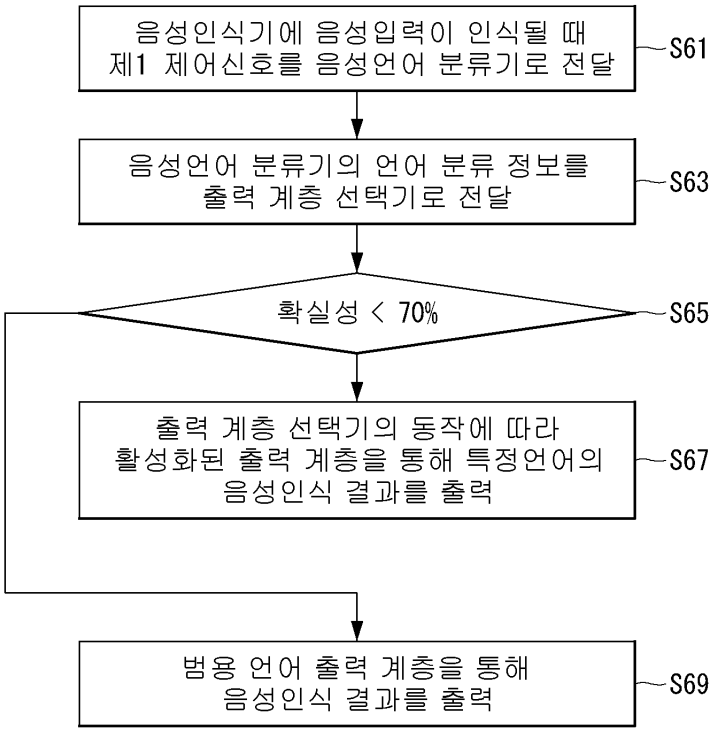
도면4


263
264
261
262
Unit (25ms)
Unit (pooled)
280

도면5


100
110
프로세서
120
메모리
입력 인터페이스 장치
150
출력 인터페이스 장치
160
송수신 장치
130
저장 장치
140

도면6


음성인식기에 음성입력이 인식될 때 제1 제어신호를 음성언어 분류기로 전달
S61
음성언어 분류기의 언어 분류 정보를 출력 계층 선택기로 전달
S63
확실성 < 70%
S65
출력 계층 선택기의 동작에 따라 활성화된 출력 계층을 통해 특정언어의 음성인식 결과를 출력
S67
범용 언어 출력 계층을 통해 음성인식 결과를 출력
S69