(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0081166
(43) 공개일자 2021년07월01일

(51) 국제특허분류(Int. Cl.)
G10L 15/00 (2006.01) *G10L 15/02* (2006.01)
G10L 15/06 (2006.01) *G10L 15/16* (2006.01)
G10L 15/183 (2013.01) *G10L 19/038* (2013.01)
G10L 25/93 (2013.01)
(52) CPC특허분류
G10L 15/005 (2013.01)
G10L 15/02 (2013.01)
(21) 출원번호 10-2019-0173437
(22) 출원일자 2019년12월23일
심사청구일자 없음

(71) 출원인
주식회사 케이티
경기도 성남시 분당구 불정로 90(정자동)
(72) 발명자
김태형
서울특별시 송파구 올림픽로35길 10, 225동 2502호 (신천동, 파크리오)
(74) 대리인
유미특허법인

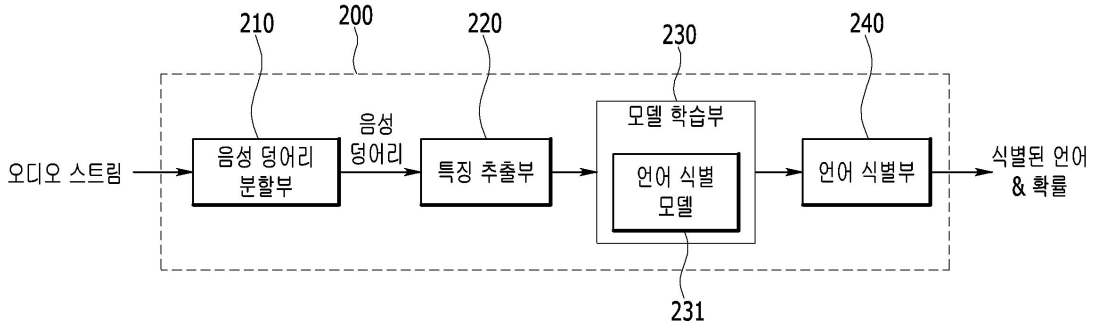
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **다국어 음성 환경에서의 언어 식별 장치 및 방법**

(57) 요 약

적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치가 언어를 식별하는 방법으로서, 오디오 스트림을 하나 이상의 음성 덩어리로 분할하는 단계, 각 음성 덩어리에서 복수의 음향 특징들을 추출하는 단계, 각 음성 덩어리에서 추출된 음향 특징들에 상기 오디오 스트림의 언어를 태깅하여 학습 데이터를 생성하는 단계, 그리고 상기 학습 데이터를 기반으로 상기 오디오 스트림의 언어를 추정하는 언어 식별 모델을 학습하는 단계를 포함하고, 상기 음성 덩어리는, 상기 오디오 스트림에서 일정 길이 이상의 음성 대역 구간을 추출한 것이다.

대 표 도



(52) CPC특허분류
G10L 15/063 (2013.01)
G10L 15/16 (2013.01)
G10L 15/183 (2013.01)
G10L 19/038 (2013.01)
G10L 25/93 (2013.01)

명 세 서

청구범위

청구항 1

적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치가 언어를 식별하는 방법으로서,

오디오 스트림을 하나 이상의 음성 덩어리로 분할하는 단계,

각 음성 덩어리에서 복수의 음향 특징들을 추출하는 단계,

각 음성 덩어리에서 추출된 음향 특징들에 상기 오디오 스트림의 언어를 태깅하여 학습 데이터를 생성하는 단계, 그리고

상기 학습 데이터를 기반으로 상기 오디오 스트림의 언어를 추정하는 언어 식별 모델을 학습하는 단계를 포함하고,

상기 음성 덩어리는, 상기 오디오 스트림에서 일정 길이 이상의 음성 대역 구간을 추출한 것인, 언어 식별 방법.

청구항 2

제1항에서,

상기 언어 식별 모델은,

학습 결과에 따라 상기 각 음성 덩어리에서 추출된 음향 특징들 각각에 서로 다른 가중치를 부여하는 순환 신경망(Recurrent Neural Networks)을 포함하는, 언어 식별 방법.

청구항 3

제1항에서,

상기 복수의 음향 특징들은,

MFCC(Mel Frequency Cepstral Coefficient)를 포함하는, 언어 식별 방법.

청구항 4

제1항에서,

상기 추출하는 단계는,

상기 오디오 스트림을 구성하는 각 음성 덩어리 마다 제1 음향 특징과 제2 음향 특징을 추출하는 단계, 그리고

추출된 제1 음향 특징들을 배열하여 제1 특징 벡터를 생성하고, 추출된 제2 음향 특징들을 배열하여 제2 특징 벡터를 생성하는 단계

를 포함하는, 언어 식별 방법.

청구항 5

제4항에서,

상기 언어 식별 모델을 학습하는 단계는,

상기 제1 특징 벡터와 상기 제2 특징 벡터를 이용하여 서로 다른 언어 식별 모델을 학습하는, 언어 식별 방법.

청구항 6

제1항에서

새로운 오디오 스트림을 하나 이상의 음성 덩어리로 분할하는 단계,

상기 새로운 오디오 스트림의 음성 덩어리 중 음향 정보를 포함한 음성 덩어리만을 상기 언어 식별 모델로 입력하는 단계, 그리고

상기 언어 식별 모델로 상기 새로운 오디오 스트림의 예측 언어를 적어도 하나 이상 추정하는 단계

를 더 포함하는, 언어 식별 방법.

청구항 7

제6항에서,

상기 입력하는 단계는,

상기 새로운 오디오 스트림에서 분할된 음성 덩어리들 중 상기 컴퓨팅 장치에 시간순으로 먼저 입력된 오디오 스트림에 해당하는 일부의 음성 덩어리만을 상기 언어 식별 모델로 입력하는, 언어 식별 방법.

청구항 8

제6항에서,

상기 추정하는 단계는,

상기 예측 언어가 복수 개인 경우, 상기 예측 언어 별 확률 값을 함께 출력하는, 언어 식별 방법.

청구항 9

컴퓨팅 장치로서,

메모리, 그리고

상기 메모리에 로드된 프로그램의 명령들(instructions)을 실행하는 적어도 하나의 프로세서를 포함하고,

상기 프로그램은

오디오 스트림에서 미리 설정된 음성 대역에 해당하는 음성 구간을 추출하고, 추출된 음성 구간의 길이가 미리 설정된 기준 길이를 초과하는 경우, 상기 음성 구간의 오디오 스트림을 음성 덩어리로 생성하는 단계,

각 음성 덩어리에서 복수의 음향 특징들을 추출하는 단계,

상기 추출된 음향 특징들을 학습된 언어 식별 모델에 입력하는 단계, 그리고

상기 언어 식별 모델로부터 상기 오디오 스트림의 예측 언어를 적어도 하나 이상 획득하는 단계

를 실행하도록 기술된 명령들을 포함하는, 컴퓨팅 장치.

청구항 10

제9항에서,

상기 추출하는 단계는,

상기 각 음성 덩어리 중 상기 컴퓨팅 장치에 시간순으로 먼저 입력된 오디오 스트림에 해당하는 일부의 음성 덩어리에서 상기 복수의 음향 특징들을 추출하는, 컴퓨팅 장치.

청구항 11

제9항에서,

상기 입력하는 단계는,

상기 추출된 음향 특징에 따라 서로 다른 순환 신경망(Recurrent Neural Networks)에 입력하는, 컴퓨팅 장치.

청구항 12

제9항에서,

상기 획득하는 단계는,

상기 오디오 스트림을 임의의 음성 인식 모델에 입력하고, 상기 음성 인식 모델로부터 음성 인식 결과를 얻는 단계, 그리고

상기 음성 인식 결과를 이용하여 상기 예측 언어 중에서 상기 오디오 스트림의 언어를 결정하는 단계

를 포함하는, 컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 다국어 음성 환경에서 언어를 식별하는 기술에 관한 것이다.

배 경 기 술

[0002] 음성 인식 기술의 발달로 마이크와 스피커를 구비한 지능형 개인비서 또는 스마트 스피커가 광범위하게 보급되고 있다. 이러한 전자 장치는 음성 대화 시스템 (Spoken Dialog System)을 사용하며, 음성 대화 시스템은 음성 인식 - 자연어 이해 - 대화 운영 - 자연어 생성 - 음성 합성의 순서로, 입력된 음성에 대응하는 음성 답변을 생성하여 출력한다.

[0003] 최근에는 이러한 전자 장치에서 활용되는 음성 대화 시스템에 딥러닝을 적용하여 음성 인식, 음성 합성의 성능을 대폭 향상시켰다. 또한 자연어 이해, 생성에서도 딥러닝을 적용하기 위한 연구가 활발히 진행되고 있다.

[0004] 한편, 음성 대화 시스템은 단일 언어에 국한되지 않고, 다수의 언어를 지원하기 시작했다. 하나의 공식 언어를 쓰는 국가에서도, 현대 사회에서 국제화 추세에 따라 다국어 자연어 인터페이스는 반드시 필요한 기술이 되었다. 예를 들면, 다양한 국적의 투숙객이 묵는 호텔 객실에서, 고객은 스마트 스피커를 통해 필요한 물품을 요청하거나, 음악을 틀거나, 컨시어지(Concierge) 서비스를 제공받을 수 있다. 또는 한 가정에서 부모와 자식의 모국어가 다른 경우에도 2개 국어로 스마트 스피커를 이용할 수 있게 되는 추세이다.

[0005] 다국어 상황에서 음성 대화 시스템은, 앞서 설명한 구성에 음성 언어 식별 (Spoken Language Identification, 이하 'LID'라고 호칭함) 과정이 추가된다. 즉, 언어 식별 - 음성 인식 - 자연어 이해 - 대화 운영 - 자연어 생성 - 음성 합성 과정으로 이루어지거나, 또는 언어 식별과 음성 인식이 병렬로 수행될 수 있다.

[0006] 다국어를 지원하는 음성 대화 시스템은 지원하는 언어를 각각 처리하는 음성 대화 서브 시스템들을 포함해야 한다. 음성 대화 시스템에서 사용자의 발화를 입력받는 채널이 하나이므로, 이러한 다국어 음성 대화 시스템의 서브 시스템들은 동일한 입력 채널을 통해 음성을 입력받게 된다. 각각의 언어를 지원하는 서브 시스템이 개별적으로 존재하므로, 앞단에서 언어 식별 과정 없이 서브 시스템이 입력된 음성을 처리하더라도 출력에서 언어를 결정해야 하는 문제가 발생한다. 따라서 서브 시스템이 모두 병렬 처리를 하는 것보다 앞단에서 언어를 식별하는 것이 시스템의 자원을 효율적으로 사용하는 방법이다.

[0007] 다국어 환경에서 음성 대화 시스템이 스스로 언어를 식별하는 기술이 적용되기 전에는 사용자가 리모컨, 스마트폰 등의 입력 장치 또는 음성 발화로 언어를 지정하였다. 예를 들어, 한국어로 설정된 전자 기기에 "영어모드로 바꿔줘" 라고 말해서 영어로 전환할 수 있다. 그리고 사용자가 기 설정된 언어에 익숙하지 않은 경우는, 목표 언어로 발화하여 언어를 전환할 수 있다. 예를 들어 영어로 설정된 전자 기기에 "한국어로 말할래" 라고 말하는 경우이다.

[0008] 최근에는 입력된 음성의 언어를 자동으로 식별하는 기술이 발달함에 따라, 사용자가 수동으로 발화 언어를 지정할 필요가 없게 되었다. 이러한 언어 식별 기술은 전통적으로 오디오 신호 처리 및 통계적 기법이 이용되었으나, 딥러닝 기법이 도입되는 추세이다.

[0009] 일반적으로 딥러닝 LID 기술은 입력 발화를 프레임 단위(예를 들어 10ms)로 나누고, 음향적 특징을 추출하여, 음성 대화 시스템이 지원하는 언어의 수에 맞게 학습된 모델로 다중 클래스 분류(Multi-Class Classification)를 수행한다. 예를 들어, 2개 국어를 지원하는 시스템은 2개의 클래스(한국어, 영어) 혹은 3개의 클래스(한국어, 영어, 기타 언어)로 분류할 수 있다. 이때, 데이터 양 또는 발화 내용과 무관하게 음향적 특징은 프레임 단위로 적용된다.

[0010] 스마트 스피커에 입력된 오디오 신호의 최소 처리 단위는 프레임이며, 스마트 스피커가 스트리밍하는 프레임을 수신하면, 음성 대화 시스템은 정보의 양을 줄이기 위해 특징을 추출하여 처리한다. 이때 스트리밍되는 프레임

에서 특징을 추출하는 횟수가 빈번하고, 학습을 위한 음성 코퍼스 데이터가 대부분 문장 단위로 구성되어 있으므로 언어 식별 정확도를 위해서는 문장이 끝날 때까지 기다려야 한다는 단점이 있다.

[0011] 이와 같이 다국어 음성 대화 시스템에 적용되는 LID 기술은 몇 가지 도전적인 환경에 놓이게 된다. 우선 기존의 LID 기술을 학습하는데 활용되던 전화 통화 데이터 등의 데이터에 비해, 스마트 스피커를 통한 입력 음성 발화의 길이가 짧은 편이다. 이에 더해 상용 시스템의 경우 사용자 경험(User Experience, UX)도 고려해야 하므로, 사용자가 발화 후 답변을 받기까지 걸리는 지연 시간을 줄이기 위한 노력이 필요하다. 그러므로 다국어 음성 대화 시스템의 한 단계인 LID도 사용자가 체감하는 지연 시간을 줄이기 위한 방법이 필요하다.

발명의 내용

해결하려는 과제

[0012] 해결하고자 하는 과제는 음성 덩어리의 특징을 이용하여 인공 신경망 기반으로 화자의 언어를 식별하는 장치와 방법을 제공하는 것이다.

[0013] 또한, 해결하고자 하는 과제는 입력되는 오디오 스트림을 가변적인 길이를 갖는 음성 덩어리로 분할하여 음성 덩어리의 특징을 추출하는 방법을 제공하는 것이다.

과제의 해결 수단

[0014] 한 실시예에 따른 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치가 언어를 식별하는 방법으로서, 오디오 스트림을 하나 이상의 음성 덩어리로 분할하는 단계, 각 음성 덩어리에서 복수의 음향 특징들을 추출하는 단계, 각 음성 덩어리에서 추출된 음향 특징들에 상기 오디오 스트림의 언어를 태깅하여 학습 데이터를 생성하는 단계, 그리고 상기 학습 데이터를 기반으로 상기 오디오 스트림의 언어를 추정하는 언어 식별 모델을 학습하는 단계를 포함하고, 상기 음성 덩어리는, 상기 오디오 스트림에서 일정 길이 이상의 음성 대역 구간을 추출한 것이다.

[0015] 상기 언어 식별 모델은, 학습 결과에 따라 상기 각 음성 덩어리에서 추출된 음향 특징들 각각에 서로 다른 가중치를 부여하는 순환 신경망(Recurrent Neural Networks)을 포함할 수 있다.

[0016] 상기 복수의 음향 특징들은, MFCC(Mel Frequency Cepstral Coefficient)를 포함할 수 있다.

[0017] 상기 추출하는 단계는, 상기 오디오 스트림을 구성하는 각 음성 덩어리 마다 제1 음향 특징과 제2 음향 특징을 추출하는 단계, 그리고 추출된 제1 음향 특징들을 배열하여 제1 특징 벡터를 생성하고, 추출된 제2 음향 특징들을 배열하여 제2 특징 벡터를 생성하는 단계를 포함할 수 있다.

[0018] 상기 언어 식별 모델을 학습하는 단계는, 상기 제1 특징 벡터와 상기 제2 특징 벡터를 이용하여 서로 다른 언어 식별 모델을 학습할 수 있다.

[0019] 새로운 오디오 스트림을 하나 이상의 음성 덩어리로 분할하는 단계, 상기 새로운 오디오 스트림의 음성 덩어리 중 음향 정보를 포함한 음성 덩어리만을 상기 언어 식별 모델로 입력하는 단계, 그리고 상기 언어 식별 모델로 상기 새로운 오디오 스트림의 예측 언어를 적어도 하나 이상 추정하는 단계를 더 포함할 수 있다.

[0020] 상기 입력하는 단계는, 상기 새로운 오디오 스트림에서 분할된 음성 덩어리들 중 상기 컴퓨팅 장치에 시간순으로 먼저 입력된 오디오 스트림에 해당하는 일부의 음성 덩어리만을 상기 언어 식별 모델로 입력할 수 있다.

[0021] 상기 추정하는 단계는, 상기 예측 언어가 복수 개인 경우, 상기 예측 언어 별 확률 값을 함께 출력할 수 있다.

[0022] 한 실시예에 따른 컴퓨팅 장치로서, 메모리, 그리고 상기 메모리에 로드된 프로그램의 명령들(instructions)을 실행하는 적어도 하나의 프로세서를 포함하고, 상기 프로그램은 오디오 스트림에서 미리 설정된 음성 대역에 해당하는 음성 구간을 추출하고, 추출된 음성 구간의 길이가 미리 설정된 기준 길이를 초과하는 경우, 상기 음성 구간의 오디오 스트림을 음성 덩어리로 생성하는 단계, 각 음성 덩어리에서 복수의 음향 특징들을 추출하는 단계, 상기 추출된 음향 특징들을 학습된 언어 식별 모델에 입력하는 단계, 그리고 상기 언어 식별 모델로부터 상기 오디오 스트림의 예측 언어를 적어도 하나 이상 획득하는 단계를 실행하도록 기술된 명령들을 포함한다.

[0023] 상기 추출하는 단계는, 상기 각 음성 덩어리 중 상기 컴퓨팅 장치에 시간순으로 먼저 입력된 오디오 스트림에 해당하는 일부의 음성 덩어리에서 상기 복수의 음향 특징들을 추출할 수 있다.

[0024] 상기 입력하는 단계는, 상기 추출된 음향 특징에 따라 서로 다른 순환 신경망(Recurrent Neural Networks)에 입

력할 수 있다.

[0025] 상기 획득하는 단계는, 상기 오디오 스트림을 임의의 음성 인식 모델에 입력하고, 상기 음성 인식 모델로부터 음성 인식 결과를 얻는 단계, 그리고 상기 음성 인식 결과를 이용하여 상기 예측 언어 중에서 상기 오디오 스트림의 언어를 결정하는 단계를 포함할 수 있다.

발명의 효과

[0026] 본 발명에 따르면, 발화된 음성의 언어를 빠르게 식별하고, 식별된 언어에 해당하는 음성 인식 서브 시스템만을 가동하여 이외의 시스템을 중단할 수 있으므로 음성 대화 시스템의 연산 자원을 절약할 수 있다.

[0027] 또한 본 발명에 따르면, 사용자의 입력 발화 전체를 기다릴 필요 없이, 먼저 입력되는 일부의 음성 덩어리를 이용하여 언어를 식별할 수 있으므로, 기존 방식보다 시스템의 처리 속도를 빠르게 할 수 있다.

도면의 간단한 설명

[0028] 도 1은 한 실시예에 따른 언어 식별 장치와 그 주변 환경의 구성도이다.

도 2는 기존의 언어 식별 장치의 구성도이다.

도 3은 한 실시예에 따른 언어 식별 장치의 구성도이다.

도 4는 한 실시예에 따른 오디오 스트림을 음성 덩어리로 분할하는 방법의 흐름도이다.

도 5는 한 실시예에 따른 음성 덩어리의 파형을 나타낸 예시도이다.

도 6은 한 실시예에 따른 음성 덩어리에서 추출된 특징으로 학습 데이터를 생성하는 방법의 설명도이다.

도 7은 다른 실시예에 따른 음성 덩어리에서 추출된 특징으로 학습 데이터를 생성하는 방법의 설명도이다.

도 8은 한 실시예에 따른 언어를 식별하는 방법의 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0029] 아래에서는 첨부한 도면을 참고로 하여 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.

[0030] 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 있는 것을 의미한다. 또한, 명세서에 기재된 "…부", "…기", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0031] 도 1은 한 실시예에 따른 언어 식별 장치와 그 주변 환경의 구성도이다.

[0032] 도 1을 참고하면, 사용자의 음성은 전자 장치(100)를 거쳐 프레임 단위의 오디오 스트림으로 가공되고, 언어 식별 장치(200)는 오디오 스트림을 음성 덩어리로 분할하고 음성 덩어리들의 특징을 추출하여 입력 벡터를 생성하고, 딥러닝을 이용한 언어 식별 모델(231)을 통해, 사용자의 언어를 특정할 수 있다.

[0033] 전자 장치(100)는 언어 식별 장치(200)와 유선 또는 무선으로 연결될 수 있으며, 사용자가 발화한 소리를 수신하는 마이크(110), 발화 음성을 디지털 신호로 변환하는 ADC(120) 그리고 음성 신호의 SNR(Signal-to-Noise Ratio)을 증폭하는 전처리부(130)를 포함한다.

[0034] 마이크(110)를 통해 입력된 사용자의 발화는 ADC(Analog-to-Digital Converter)(120)를 통해 디지털 신호로 변환된다.

[0035] ADC(120)는 입력되는 아날로그 신호를 특정 샘플링 레이트(Sampling Rate)와 특정 비트 깊이(Bit Depth), 예를 들어 샘플링 레이트를 16 kHz로, 비트 깊이를 16 bit로 설정하여 디지털 신호로 변환할 수 있다.

[0036] 전처리부(130)는 디지털 신호 중 발화 신호(Signal)는 향상시키고, 잡음(Noise)은 억제한다. 신호 처리는 시간

영역(Time Domain) 혹은 주파수 영역(Frequency Domain)에서 이루어질 수 있으며, 노이즈 취소(Noise Cancellation), 에코 취소(Echo Cancellation), 목소리 활동 감지(Voice Activity Detection) 등이 적용될 수 있다.

[0037] 도 2는 기존의 언어 식별 장치의 구성도이다.

[0038] 도 2를 참고하면, 기존의 언어 식별 장치(20)에서는, 실시간으로 입력되는 오디오 스트림은 사용자로부터 순차적으로 입력되는 오디오 샘플의 프레임들을 의미한다. 하나의 문장을 발화한 파형(Waveform)의 전체 샘플을 시각화하면, 프레임 길이의 윈도우를 홉(Hop) 길이만큼 건너뛰는 것을 입력 신호의 처음부터 끝까지 수행한 것과 같다. 이 때 붙어있는 프레임들은 앞 프레임과 뒷 프레임이 오버랩(Overlap) 구간만큼 겹치게 된다. 예를 들어, 10 ms 길이의 프레임에서 홉의 길이를 2 ms로 설정하면 다음 프레임과의 오버랩 구간이 8 ms가 된다. 발화의 시작의 묵음 구간과 발화가 끝날 때의 묵음 구간을 포함한 총 발화 시간이 0.91 초인 경우를 가정할 때, 0.91 초에서 10 ms를 제외하고 2 ms 단위로 프레임을 분할하면 총 프레임의 개수는 450개이다.

[0039] 이때 기존 장치(20)는 450개의 프레임에 대해 각 프레임마다 특징을 추출해야 하므로 시스템 자원의 낭비를 초래할 수 있다. 또한 언어를 식별하기 위해 문장이 끝나거나 끝점 인식(End Point Detection)이 되는 것을 기다려야 하므로 시스템 처리 시간의 지연이 발생할 수 있다.

[0040] 이렇게 분할된 각각의 프레임에 대해 특징 추출부(21)에서 음향 특징을 추출한다. 특징 추출부(21)가 음향 특징을 추출하는 방법은 인간의 청각 기관을 연구한 결과를 이용할 수 있다. 음향 특징(Acoustic Feature), 운율 특징(Prosodic Feature), 음소 특징(Phonotactic Feature) 등이 적용될 수 있으며, 경우에 따라 자연어에 대한 이해를 바탕으로 추가적으로 어휘 특징(Lexical Feature), 구문 특징(Syntactic Feature)을 활용할 수 있다. 대표적인 음향 레벨 특징으로는 MFCC(Mel-Frequency Cepstral Coefficient), SDC(Shifted Delta Cepstral coefficient) 등이 있다.

[0041] 추출한 특징들을 기반으로 모델 학습부(22)는 언어 식별 모델(23)을 학습하게 된다. 예를 들어 MFCC 또는 SDC를 이용하여 N개의 가우시안 요소(Gaussian Components)를 이용한 GMM(Gaussian Mixture Model)을 생성할 수 있다. 또는 사용자가 데이터와 특징을 입력하고 이에 상응하는 라벨을 이용하여 모델을 학습하는 머신러닝 또는 딥러닝 알고리즘으로 구현될 수 있다.

[0042] 이후, 언어 식별부(24)는 학습된 언어 식별 모델(23)을 이용하여 프레임 단위로 입력되는 오디오 스트림의 발화 언어를 결정한다.

[0043] 즉 기존 언어 식별 장치(20)는 프레임 단위로 특징을 추출하는데, 본 명세서에서 제안하는 발명은 프레임보다 큰 단위인 소리 덩어리 단위로 음향 특징을 추출하는 것이 가장 큰 특징이다. 이하에서는, 본 명세서에서 제안하는 언어 식별 장치(200) 및 동작 방법에 대해 설명한다.

[0044] 도 3은 한 실시예에 따른 언어 식별 장치의 구성도이다.

[0045] 도 3을 참고하면, 언어 식별 장치(200)는 오디오 스트림을 음성 덩어리로 분할하는 음성 덩어리 분할부(210), 음성 덩어리의 특징을 추출하고 입력 벡터(410)를 생성하는 특징 추출부(220), 입력 벡터(410)를 이용하여 언어 식별 모델(231)을 학습하는 모델 학습부(230) 그리고 언어 식별 모델(231)이 생성한 기준에 따라 입력 발화의 언어를 판별하고, 해당 언어일 확률을 출력하는 언어 식별부(240)를 포함한다.

[0046] 설명을 위해, 음성 덩어리 분할부(210), 특징 추출부(220), 모델 학습부(230) 그리고 언어 식별부(240)로 명명하여 부르나, 이들은 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치이다. 여기서, 음성 덩어리 분할부(210), 특징 추출부(220), 모델 학습부(230) 그리고 언어 식별부(240)는 하나의 컴퓨팅 장치에 구현되거나, 별도의 컴퓨팅 장치에 분산 구현될 수 있다. 별도의 컴퓨팅 장치에 분산 구현된 경우, 음성 덩어리 분할부(210), 특징 추출부(220), 모델 학습부(230) 그리고 언어 식별부(240)는 통신 인터페이스를 통해 서로 통신할 수 있다. 컴퓨팅 장치는 본 발명을 수행하도록 작성된 소프트웨어 프로그램을 실행할 수 있는 장치이면 충분하고, 예를 들면, 서버, 랩탑 컴퓨터 등일 수 있다.

[0047] 특징 추출부(220), 모델 학습부(230) 그리고 언어 식별부(240) 각각은 하나의 딥러닝 모델일 수 있고, 복수의 딥러닝 모델로 구현될 수도 있다. 그리고 언어 식별 모델(231)도 하나의 딥러닝 모델일 수 있고, 복수의 딥러닝 모델로 구현될 수도 있다. 언어 식별 장치(200)는 하나의 딥러닝 모델일 수 있고, 복수의 딥러닝 모델로 구현될 수도 있다. 이에 따라, 상술한 구성들에 대응하는 하나 또는 복수의 딥러닝 모델은 하나 또는 복수의 컴퓨팅 장치에 의해 구현될 수 있다.

[0048] 음성 덩어리 분할부(210)는 오디오 스트림으로 입력되는 프레임이 음성(Speech)인지를 판단하는 알고리즘에 따라 음성 덩어리를 형성한다. 구체적으로, 하나의 음성 덩어리가 있는 예를 통해 설명한다. 첫 번째 프레임부터 n번째 프레임까지 음성이 아니고, n+1번째 프레임부터 m개의 프레임이 음성이고, n+m+2번째 프레임부터 마지막 프레임까지 음성이 아닌 것으로 판단되는 경우, 음성 덩어리 분할부(210)는 음성으로 판단된 m개의 프레임을 모아서 음성 덩어리를 생성할 수 있다.

[0049] 한편, 음성으로 판단된 구간이 복수일 경우, 복수의 음성 덩어리가 생성된다. 음성 덩어리 분할부(210)는 음성으로 판단된 연속적인 프레임으로 구성된 음성 덩어리들을 다음 단인 특징 추출부(220)로 전달한다.

[0050] 특징 추출부(220)는 각 음성 덩어리의 음향 특징을 추출하여 입력 문장에 대한 입력 벡터(410)를 생성한다. 특징 추출부(220)는 사용자 또는 개발자가 특정한 시간, 주파수 상의 다양한 특징들을 추출할 수 있다. 다른 방법으로서, 특징 추출부(220)는 인공 신경망의 구조에 따라 학습 과정에서 임의의 특징을 자동으로 추출할 수도 있다. 특징 추출부(220)에서 각 덩어리마다 추출된 음향 특징은 입력 벡터(410)로 생성되어 학습 데이터로서 사용된다.

[0051] 모델 학습부(230)는 입력 벡터(410)와 해당하는 언어를 판단한 라벨링 결과를 이용하여 언어 식별 모델(231)을 학습한다. 본 발명에서는 데이터가 음성 덩어리 단위로 구분되므로 음성 덩어리마다 라벨링 결과가 다를 수 있다.

[0052] 본 명세서에서 언어 식별 모델(231)은 순환 신경망(RNN)으로 구현되나, 모델을 구현하는 딥러닝 알고리즘은 반드시 이에 한정되는 것은 아니다.

[0053] 언어 식별부(240)는 학습된 언어 식별 모델(231)을 이용하여, 사용자 발화의 언어를 결정하고, 결정된 언어일 확률 값을 함께 출력할 수 있다. 언어를 결정할 때, 결정이론(Decision Theory)을 바탕으로 다양한 방법이 활용될 수 있다.

[0054] 한 예로서, 계산된 확률 값이 가장 높은 언어 하나를 선택할 수 있다. 다른 예로서, 확률 값이 미리 정한 기준 값을 넘는 언어를 출력할 수 있으며, 이때 언어 식별부(240)가 결정한 언어는 복수일 수 있다.

[0055] 또 다른 예로서, 동일한 입력 신호에 대해 언어 식별부(240)가 언어 식별 모델(231)을 이용하여 언어를 결정하는 과정과, 각 언어의 음성 대화 서브 시스템의 음성 인식 과정이 병렬로 진행될 수 있다. 이때 언어 식별부(240)는 언어 식별 모델(231)을 이용하여 계산한 확률 값과 각 언어의 음성 인식 결과에서 추출한 신뢰도 등의 특정값을 입력으로 하는 또 다른 분류기를 사용하여 언어를 결정할 수 있다. 즉, 사용자 발화에 대해 언어 식별과 음성 인식을 동시에 수행하고, 학습된 모델을 이용한 결과와 음성 인식 결과를 고려하여 해당 발화의 언어를 결정할 수 있다.

[0056] 또한, 언어 식별과 음성 인식이 병렬로 진행되면 언어 식별부(240)의 판단이 음성 인식 과정보다 빠를 경우, 식별되지 않은 언어 또는 식별될 확률이 현저하게 낮은 언어의 음성 인식 과정을 중단시키므로 시스템의 자원 낭비를 막을 수 있다.

[0057] 한편 언어 식별부(240)가 출력하는 결과와, 언어를 결정하는 방법은 언어 식별 장치(200)의 활용 방안, 정책, 전략 등에 따라 변할 수 있고 어느 하나에 한정되지 않는다.

[0058] 도 4는 한 실시예에 따른 오디오 스트림을 음성 덩어리로 분할하는 방법의 흐름도이다.

[0059] 도 4를 참고하면 음성 덩어리 분할부(210)는 프레임 카운트와 버퍼를 리셋한다(S101). 프레임 카운트는 입력되는 오디오 스트림의 길이를 측정하기 위해 사용되며, 버퍼는 음성으로 판별된 프레임을 임시 저장하는 역할을 한다.

[0060] 음성 덩어리 분할부(210)는 프레임 단위의 오디오 스트림을 수집한다(S102).

[0061] 이후 음성 덩어리 분할부(210)는 입력된 오디오 스트림에 가우시안 필터를 이용하여 고주파 성분을 감쇄시킨다(S103). 또한 가우시안 필터를 사용하면 그래프 상에서 음성인 신호와 음성이 아닌 신호 사이의 골(Valley)이 깊어지므로 입력된 오디오 스트림이 음성인지를 판단하는데 도움이 될 수 있다.

[0062] 음성 덩어리 분할부(210)는 수집된 현재 프레임이 음성(Speech)인지 여부를 확인한다(S104). 음성인지 여부를 확인하는 방법 중 한 예로서, 음성 활동 감지(Voice Activity Detection, 이하 'VAD'라고 호칭함) 기술을 사용할 수 있다. VAD 기술을 적용하면 음성 신호의 시간 및 주파수 분석을 통해 입력된 신호 구간이 사람의 음성을

포함하는지 여부를 판단할 수 있다.

[0063] 이하, VAD 기술에 대해 상세히 설명한다. VAD 기술은 음성 부호화(Speech Coding), 음성 인식(Speech Recognition), 음성 개선(Speech Enhancement) 등 대부분의 음성 처리(Speech Processing) 분야에서 널리 쓰이는 기저 기술이다. 기술적으로 크게 에너지 임계값 설정(Energy Thresholding), 음성 특징을 추가한 감지 기법(Detection with More Features), 에너지 및 음성 특징을 활용한 통계적 기법, 음성/비음성으로 분류하는 머신 러닝 혹은 딥러닝 기법으로 구분될 수 있으며, G.729 Annex B의 VAD 모듈, WebRTC(Real-Time Communication) VAD 등으로 알려져 있다. 또한 VAD 기술은 Speech Activity Detection 혹은 Speech Detection로도 호칭될 수 있다.

[0064] 한편, 다른 예로서 수집된 현재 프레임이 음성(Speech)인지 여부를 확인하기 위해 음성 존재 확률(Speech Presence Probability, 이하 'SPP'라고 호칭함)을 활용할 수 있다. 이하에서는 VAD 기술과 연관하여 SPP에 대해 상세히 설명한다.

[0065] SPP 추정(SPP Estimation)은 VAD와 밀접히 연관된 기술로서, VAD의 출력값이 SPP 추정을 통해 음성/비음성 여부를 판단한 결과를 의미한다. 즉, SPP 추정 결과는 0에서 1사이 값인 확률이고, VAD 결과는 음성인지 여부를 참/거짓으로 나타낸 것이다. 따라서 SPP 추정 결과에 임계값을 설정하여 VAD 결과를 도출할 수 있다. 예를 들어 음성신호의 프레임이 가지는 에너지로 SPP를 계산하고 미리 설정한 임계값보다 높을 경우 해당 프레임을 음성으로 판단하도록 할 수 있다. 한편, 임계값은 나누어지는 음성 덩어리의 개수가 입력된 문장에 포함된 단어의 개수에 따라 달라질 수 있다.

[0066] 다시 도 4로 돌아와서, 음성 덩어리 분할부(210)는 수신한 현재 프레임이 음성으로 판단되면, 현재 프레임을 버퍼에 저장하고, 프레임 카운트를 1 증가시킨다(S105). 그리고, 다시 S102 단계로 돌아가 새로운 프레임을 수집한다.

[0067] 음성 덩어리 분할부(210)는 입력된 현재 프레임이 음성이 아니라고 판단되면, 이전 프레임이 음성인지 판단한다(S106). 이전 프레임의 음성 여부를 판단하기 위해 S104 단계에서 설명된 기술이 사용될 수 있다.

[0068] S106 단계에서 이전 프레임이 음성이 아니라고 판단되면, 유의미한 음성 구간이 끝난 것으로 판단한다. 음성 덩어리 분할부(210)는 버퍼에 저장된 프레임의 수를 세는 프레임 카운트를 리셋하고(S107), 다시 S102 단계로 돌아가 프레임 수집을 진행한다.

[0069] S106 단계에서 이전 프레임이 음성인 경우, 음성 덩어리 분할부(210)는 현재 프레임부터 유의미한 음성 구간이 끝난 것으로 판단하고, 버퍼를 저장해서 음성 덩어리를 생성할 준비를 한다. 우선 음성 덩어리 분할부(210)는 프레임 카운트가 기준값을 초과하는지 확인한다(S108). 프레임 카운트는 곧 버퍼에 저장된 프레임의 길이를 의미하므로 버퍼에 임시 저장된 내용이 미리 설정해 놓은 최소 음성 덩어리 사이즈보다 큰 경우에만 음성 덩어리를 생성한다. 이는 무의미하게 길이가 짧은 다수의 음성 덩어리가 발생하는 것을 방지하기 위함이다. 한편, 기준값은 관리자 또는 언어 식별 장치(200)에 의해 설정되며, 변경될 수 있는 값이다.

[0070] 프레임 카운트가 기준값을 초과하는 경우, 음성 덩어리 분할부(210)는 버퍼에 저장된 프레임들을 하나의 음성 덩어리로 저장한다(S109).

[0071] 도 5는 한 실시예에 따른 음성 덩어리의 파형을 나타낸 예시도이다.

[0072] 도 5를 참고하면 사용자가 "Check me out."이라는 문장을 발화하였고, 음성이 프레임 단위로 스트리밍 된 상황을 예로 들어 설명한다. 이때, 동일한 문장에 대해 서로 다른 사용자들의 발화가 입력되는 경우, 사용자의 발화 성량, 습관 등의 요소에 따라 입력 신호의 파형이 달라지므로 문장이 동일하여도 사용자 음성의 파형(Waveform)이 달라질 수 있고, 생성된 음성 덩어리의 길이는 서로 다를 수 있다.

[0073] 310은 사용자 음성의 파형이고, 파형의 시작은 호출어가 인식된 지점이고 끝점은 끝점 검출(End Point Detection, EPD) 알고리즘을 이용하여 얻은 지점일 수 있다.

[0074] 320은 오디오 파형에 각 단어가 해당되는 범위를 구분해 놓은 그림으로서 각 단어와 소리의 덩어리가 구분되지 않음을 알 수 있다. 파형상 왼쪽에서 세번째 덩어리는 ‘Check’의 ‘k’에서 시작하여, ‘me’를 포함하고, ‘out’의 ‘ou-‘로 끝나는 소리의 파형이다.

[0075] 따라서 파형으로만 볼 때, ‘k’와 ‘me’가 연결되는 부분이 어디인지, 또는 ‘me’가 끝나고 ‘ou-‘이 시작되는 부분이 어디인지 정확히 찾아내기는 쉽지 않다.

[0076] 310과 같이 단어끼리 연결된 파형을 나타내는 오디오 스트림을 단어 단위로 정확히 분리하기 위해 음성 분할(Speech Segmentation) 기술이 사용된다. 음성 분할 기술은 단어(Word), 음절(Syllable), 음소(Phoneme) 등의 경계를 분절하는 기술로서, 음성 인식, 음성 합성 등의 음성 분석 문제의 기반 기술로 사용된다.

[0077] 330은 음성 정보를 가지지 않는 지점을 기준으로 파형을 잘라내어, 음성 덩어리를 표시한 것이다. 331은 ‘Chec-‘, 332는 ‘-k me ou-‘, 그리고 333은 ‘-t’를 나타내는 것을 파형을 통해 알 수 있다.

[0078] 340은 음성 덩어리 분할부(210)가 사용자의 발화를 분할한 결과를 나타낸 것이다. 343은 도 4의 S106 단계에서 기준 길이에 못 미치므로, 음성 덩어리 분할부(210)에서는 유효한 음성 덩어리로 취급하지 않는다. 따라서 341, 342 그리고 344만이 유효한 음성 덩어리로 생성된다.

[0079] 이상적인 경우(330)와 비교하면 342의 '-u-'가 제거된 것을 볼 수 있다. 그러나 341, 342, 344를 연속하여 재생하면 원래의 문장과 매우 유사하게 들리며, 사람이 언어를 판별하고 심지어는 내용을 이해하는데 지장을 주지 않는다. 따라서 정보 손실이 크지 않으며 동시에 음성 덩어리 분할부(210)에 의해 연산량이 조금 낮아질 수 있다.

[0080] 한편, 동일한 문장에 대해, 화자가 달라지는 경우 화자의 성량, 발음 특성 등의 차이로 인해 생성된 음성 덩어리의 길이 또는 개수가 달라질 수 있다. 예를 들어, “Please call”이라는 문장을 2 명의 화자가 발음하는 경우 화자의 발화 방식에 따라 음성 덩어리는 “Please” 또는 “Please c-” 일 수 있다.

[0081] 도 6은 한 실시예에 따른 학습 데이터를 생성하는 방법의 설명도이다.

[0082] 도 6을 참고하면, 언어 식별 모델(231)의 학습 데이터는 문장을 구성하는 각 음성 덩어리(341, 342, 344)에서 음향 특징이 추출되고, 문장 순서에 맞추어 배열된 입력 벡터(410)이다. 한편, 도 4의 기준에 따라 음성이 아닌 구간(343)은 음성 덩어리로 분할되지 않으므로, 3개의 음성 덩어리(341, 342, 344)에 대해서만 특징이 추출된다.

[0083] 특징 추출부(220)는 각 덩어리에서 특징들을 추출하여 입력 신호에 대해 하나의 입력 벡터(410)를 생성하여 입력 신호의 차원을 대폭 축소한다.

[0084] 예를 들어, 각 음성 덩어리마다 MFCC(Mel Frequency Cepstral Coefficient)와 같은 음향 특징들을 추출할 수 있다. MFCC를 계산하기 위해서는 입력된 소리를 일정 구간으로 나누고, 해당 구간마다 파워 스펙트럼(Power Spectrum)을 계산한다. 이후 파워 스펙트럼에 Mel 필터 뱅크를 적용하고, 각 필터에 에너지를 합한다. 그리고 모든 필터 뱅크 에너지의 로그값에 이산 코사인 변환(Discrete Cosine Transform, DCT)를 취하고, 일부의 계수들을 이용한다. MFCC는 이미 알려진 기술이므로 자세한 설명은 생략한다.

[0085] 한편, 특징 추출부(220)가 각 음성 덩어리에서 음향 특징을 추출하는 방법 및 추출되는 음향 특징은 이에 한정되지 않는다. 특징 추출부(220)는 음의 지속 시간, 음고(Pitch) 등 언어 식별에 적용 가능한 다양한 음향, 운율, 음소, 어휘적인 특징 등 시간 및 주파수 상의 다양한 특징들을 추출할 수 있다. 또한, 시계열 데이터 분석에 효과적인 웨이블릿 산란(Wavelet Scattering) 기법을 적용할 수도 있다.

[0086] 특징 추출부(220)는 추출한 음성 덩어리 각각의 특징을 발화 문장 순서에 맞게 벡터화할 수 있다. 음성 덩어리가 배열되는 순서에 따라 입력 벡터(410)의 값이달라질 수 있으나, 입력 벡터(410)가 배열되는 순서가 달라짐에 따라 벡터를 형성하는 축이 같이 변화하므로 순서는 중요하지 않을 수 있다. 본 발명에서는 결과를 해석할 때의 편의상 음성 덩어리가 들어오는 순서에 따라 벡터화하는 것으로 설명한다.

[0087] 특징 추출부(220)는 생성된 입력 벡터(410)를 모델 학습부(230)에 전달한다.

[0088] 이후 모델 학습부(230)는 생성된 입력 벡터(410)와 해당 언어의 라벨을 언어 식별 모델(231)에 입력하여 학습한다. 언어 식별 모델(231)을 지도 학습(Supervised Learning) 방식으로 훈련하기 위한 라벨은 훈련 결과를 지도할 신호(Supervised Signal)를 의미하며, 지도 신호, 목표 신호(Target Signal), 정답(Answer), 태그(Tag)로 호칭될 수 있다.

[0089] 한편 언어 식별 모델(231)은 다양한 구조의 딥러닝 모델로서 구현될 수 있는데, 예를 들어 순환 신경망(RNN)의 한 종류인 다층 구조 양방향 LSTM (Multi-Layer Bidirectional Long-Short Term Memory)에 다층 퍼셉트론(Multi-Layer Perceptron, MLP)이 연결된 구조가 사용될 수 있다. 이하에서는 인공 신경망에 대한 일반적인 설명을 추가한다.

[0090] 언어 식별에 일반적으로 널리 쓰이는 인공 신경망으로는 다층 퍼셉트론(Multi-Layer Perceptron, MLP), 합성곱

신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN) 등이 있다.

[0091] MLP는 FFNN(Feed-Forward Neural Network) 또는 FCN(Fully Connected Network) 또는 DN(Dense Network)으로 불리기도 하는데, 입력 데이터를 문제 공간(Problem Space)에 매핑해서, 입력 벡터가 형성한 클러스터를 분류하는데 쓰일 수 있다. 언어 식별 문제에서 N개의 언어는 N개의 클래스(Class)로 구분될 수 있다.

[0092] 분류하는 과정은, 하나의 클러스터가 하나의 클래스로 구분될 수 있도록, MLP의 가중치들(Weights)이 형성하는 비선형적인 초평면(Hyper Plane)이 결정 경계 (Decision Boundary)(520)의 역할을 하도록 가중치들을 학습하는 것이다. 가중치들을 학습하는 방법은 다양하며, 한 예로서 역전파 알고리즘(Backpropagation Algorithm) 계열의 알고리즘이 사용될 수있다.

[0093] 즉, 인공 신경망을 이용한 언어 식별 과정은, N개의 언어로 라벨링 된 학습 데이터가 문제 공간에서 클러스터를 형성하도록 하고, 이 클러스터들을 N개의 클래스로 나눌 수 있도록 결정 경계(520)를 학습하여, 이 결정 경계(520)를 바탕으로 클래스를 결정하게 된다.

[0094] 학습 과정에서 인공 신경망의 성능을 향상시켜 통계적으로 올바른 결정을 할 수 있도록 최적의 초매개변수(Hyperparameter)를 선택하게 된다. 초매개변수란 임의의 모델을 학습시킬 때, 사전에 설정하는 변수를 의미하며, 데이터에 대해 라벨을 부호화(Encoding)하는 방식, 인공 신경망의 구조 및 내부의 활성 함수(Activation Function), 손실 함수(Loss Function), 학습 알고리즘 관련 계수, 결정 기준(Decision Criterion), 성능 측정 기준(Performance Metric) 등이 있다.

[0095] 다른 인공 신경망 구조로, CNN은 FCN의 밑단에 특징을 추출하는 컨볼루션 망(Convolution Network)을 추가적으로 가진다. CNN은 FCN이 데이터를 분류할 때 필요한 특징을 컨볼루션 망이 자동으로 학습하므로, 복잡한 공간에 대한 정보를 처리하는 능력에 강점이 있다.

[0096] 또 다른 인공 신경망 구조로, RNN을 사용할 수 있다. MLP에서 입력 데이터는 모든 노드를 한번씩 지나가서 출력되는 것과 달리, RNN은 은닉층의 결과가 다시 같은 은닉층의 입력으로 들어가도록 유닛 간의 연결이 순환적 구조를 갖는다.

[0097] 즉, RNN은 현재까지 계산된 결과를 내부의 메모리(Hidden State)에 저장하고 있어, 필기체 인식이나 음성 인식 등 시공간적(Spatiotemporal) 특징을 가지는 데이터를 처리하기에 적합한 모델이다. RNN의 예로서 장단기 메모리(Long-Short Term Memory, LSTM), 게이트 순환 유닛(Gated Recurrent Unit, GRU) 등이 널리 쓰이나 하나의 구조에 한정되지 않으며 학습이 가능한 가중치에 피드백이 있는 구조라면 어떤 알고리즘이든 적용될 수 있다. RNN을 형성하는 구조에서 하나의 유닛인 LSTM, GRU 등을 단일층으로 하거나 계층적으로 쌓은 구조(Stacked Architecture)를 적용하거나, 데이터의 흐름을 단방향 혹은 양방향으로 하는 등의 변형을 할 수 있음은 물론이다.

[0098] 또한 RNN은 연속적인 음성 덩어리의 흐름, 즉 이전에 들어온 음성 덩어리의 특징을 고려할 수 있으므로, 언어 식별 모델(231) 학습 시 입력 문장에서 언어 식별에 유용한 정보가 포함된 특정 음성 덩어리를 분석하는데 유리하다.

[0099] 한편, 한 문장에 복수의 언어가 포함된 경우, 음성 덩어리의 전후를 고려해야 한다. 예를 들어 "체크인 해줘." 라는 문장이 입력되어, "체크인"과 "해줘"라는 음성 덩어리 2개로 나뉜 경우에 대해 설명한다. 첫 번째 덩어리만으로는 한국어 "체크인"인지 영어 "Check in"인지 분간하기 힘들다. 두 번째 덩어리인 "해줘"가 있는 경우, 해당 문장이 한국어라고 판단할 수 있다.

[0100] 다시 도 6으로 돌아가서, 모델 학습부(230)는 음성 덩어리에서 추출된 특징으로 구성된 벡터를 입력으로 받는다. 도 6을 참고하면, 3개의 음성 덩어리에서 추출된 특징을 모두 합친 벡터가 모델 학습부(230)에 입력된다.

[0101] 모델 학습부(230)는 순서대로 지금까지 모인 음성 덩어리의 추출된 특징으로 구성된 입력 벡터를 처리할 수 있다. 첫 음성 덩어리(341)에 대해 추출된 특징이 입력 벡터로 취급되고, 두번째 음성 덩어리(342)에 대한 특징(412)이 추출되면 앞서 추출된 특징(411)과 새로운 특징(412)이 연결된 것으로 입력 벡터를 형성하고, 마지막 음성 덩어리(344)에 대한 특징(413)이 추출될 때, 3개의 특징(411, 412, 413)을 모두 연결해서 입력 벡터를 형성한다. 따라서 마지막에 생성되는 벡터는 도 6의 입력 벡터(410) 전체와 동일하다.

[0102] 한편, 모델 학습부(230)는 입력된 문장 전체에 입력 벡터를 형성하지 않을 수 있다. 즉 먼저 들어온 음성 덩어리(341)로만 입력 벡터를 형성할 수 있고, 2개의 음성 덩어리(341, 342)로만 입력 벡터를 형성할 수 있다. 음성

덩어리 일부만을 입력 벡터로 형성하여 이후의 단계를 진행하는 경우, 전체 발화를 입력으로 하여 처리하는 것에 비해 빠른 처리가 가능하다. 이때 먼저 형성된 음성 덩어리 중 몇개를 사용하여 입력 벡터를 형성할 것인지는 관리자에 의해 결정될 수 있다.

[0103] 다른 방법으로서, 모델 학습부(230)는 자연어 처리의 n-gram 방식처럼 전체가 아닌 n개의 음성 덩어리만을 사용할 수 있다. unigram 방식처럼 현재 음성 덩어리만으로 입력 벡터를 형성하거나, bi-gram 방식처럼 현재와 이전의 음성 덩어리 2개로 입력 벡터를 형성한다. 이에 따르면, 첫 음성 덩어리(341)로 입력 벡터를 형성하고, 첫 번째와 두 번째 음성 덩어리(341, 342)로 다음 입력 벡터를 형성하고, 두 번째와 세 번째 음성 덩어리(342, 344)로 입력 벡터를 형성할 수 있다.

[0104] 한편, 한 가지 종류의 특징을 추출한 경우와 달리, 음성 덩어리에서 추출하는 음향 특징의 종류가 복수 개일 수 있으며 이하 도 7을 통해 설명한다.

[0105] 도 7은 다른 실시예에 따른 음성 덩어리에서 추출된 특징으로 학습 데이터를 생성하는 방법의 설명도이다.

[0106] 도 7을 참고하면, 음성 덩어리 분할부(210)가 입력되는 오디오 스트림을 3개의 음성 덩어리(341, 342, 344)로 생성하면, 특징 추출부(220)는 각 음성 덩어리마다 2개씩의 특징을 추출하여 2개의 벡터(410, 410')를 생성할 수 있다. 또한 추출한 음향 특징의 개수만큼의 별도의 네트워크(미도시)를 사용할 수 있다. 2개의 네트워크는 각각 순서대로 입력되는 음성 덩어리의 특징을 누적한 입력 벡터(410, 410')를 각각 학습할 수 있다. 독립된 2개의 네트워크는 3개의 음성 덩어리(341, 342, 344)로부터 추출된 각 특징들의 잠재된 의미를 독립적으로 학습할 수 있다. 이후 모델 학습부(230)는 2개의 네트워크에서 각각 출력된 벡터들을 이어서 잠재 벡터를 형성하고, 잠재 벡터는 다음 단계의 네트워크인 MLP와 같은 분류기로 입력될 수 있다.

[0107] 도 8은 한 실시예에 따른 언어를 식별하는 방법의 예시도이다.

[0108] 도 8을 참고하면 언어 식별 모델(231)은 도 4의 특징 추출부(220)에서 생성된 입력 벡터(410)를 입력으로 하고, N개의 언어에 대한 확률 값을 출력한다. 또한 언어 식별 모델(231)의 학습 결과는, 벡터 공간(510)에 생성된 결정 경계(520)로 표현될 수 있다.

[0109] 즉, 모델 학습부(230)는 입력 벡터(410)들의 벡터 공간(510)을 형성하고, 벡터 공간(510) 상에 결정 경계(520)를 형성하여 각 언어에 해당하는 공간을 분류한다.

[0110] 모델 학습부(230)가 비선형적인 결정 경계(520)를 형성하는 방법은 통계적, 머신러닝, 딥러닝일 수 있으며, 대표적인 머신러닝 기법의 예로서 서포트 벡터 머신(Support Vector Machine, SVM)을 사용할 수 있다. SVM의 동작 방법은 널리 알려진 것이므로 자세한 설명은 생략한다. 또한 대표적인 딥러닝 기법의 예로서 MLP, CNN, RNN 등을 사용할 수 있으며, 동작 방법은 앞서 설명한 바와 같다.

[0111] 결정 경계(520)가 학습된 이후, 새로운 입력 벡터가 입력되면, 해당 벡터는 결정 경계(520)에 따라 도 8과 같이 언어 1로 결정될 수 있다. 또한 언어 식별부(240)는 새로운 입력 벡터가 언어 1에 해당할 확률 값을 함께 출력할 수 있다.

[0112] 본 발명에 따르면, 발화된 음성의 언어를 빠르게 식별하고, 식별된 언어에 해당하는 음성 인식 서브 시스템만을 가동하여 이외의 시스템을 중단할 수 있으므로 음성 대화 시스템의 연산 자원을 절약할 수 있다.

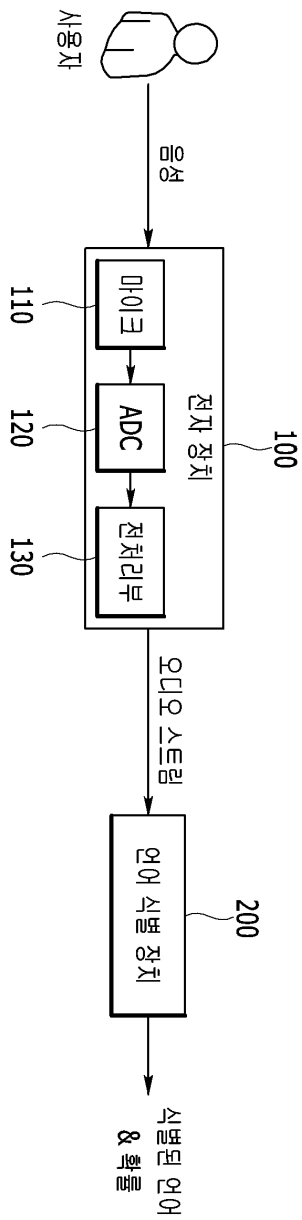
[0113] 또한 본 발명에 따르면, 사용자의 입력 발화 전체를 기다릴 필요 없이, 먼저 입력되는 일부의 음성 덩어리를 이용하여 언어를 식별할 수 있으므로, 기존 방식보다 시스템의 처리 속도를 빠르게 할 수 있다.

[0114] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.

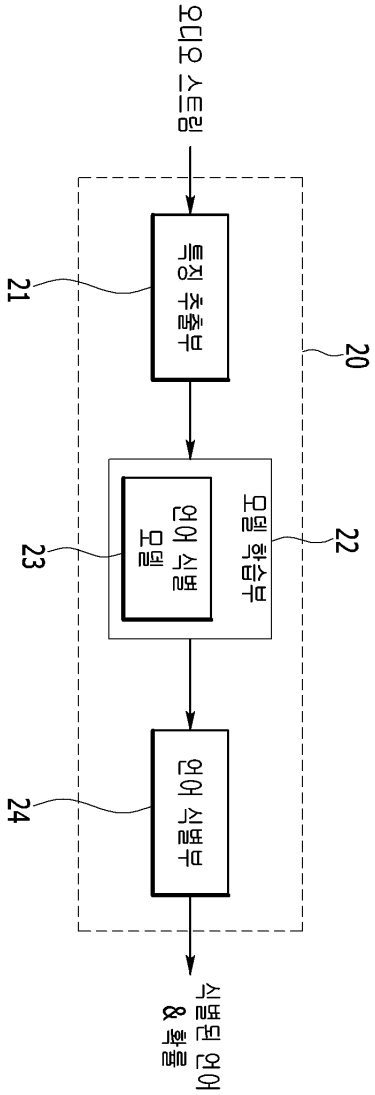
[0115] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

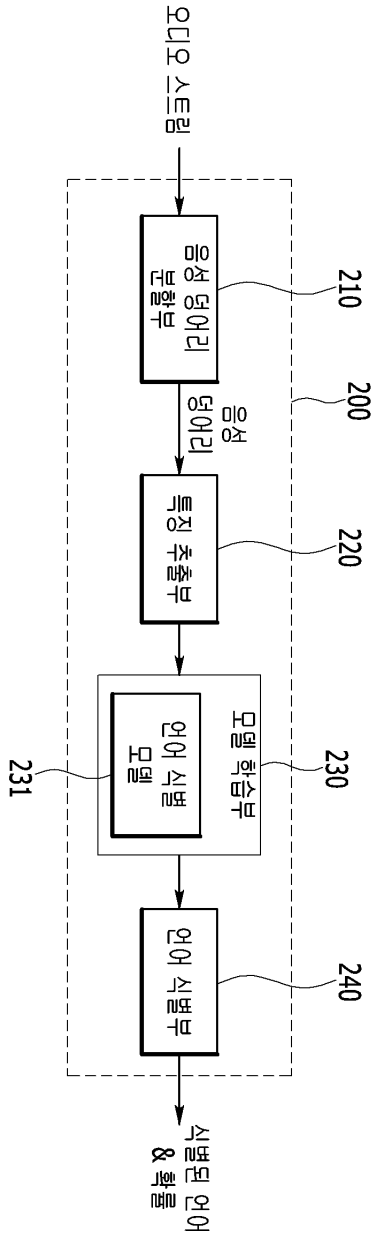
도면1


사용자
음성
마이크
110
ADC
120
전처리부
130
전자 장치
100
오디오 스트림
언어 식별 장치
200
식별된 언어
&
확률

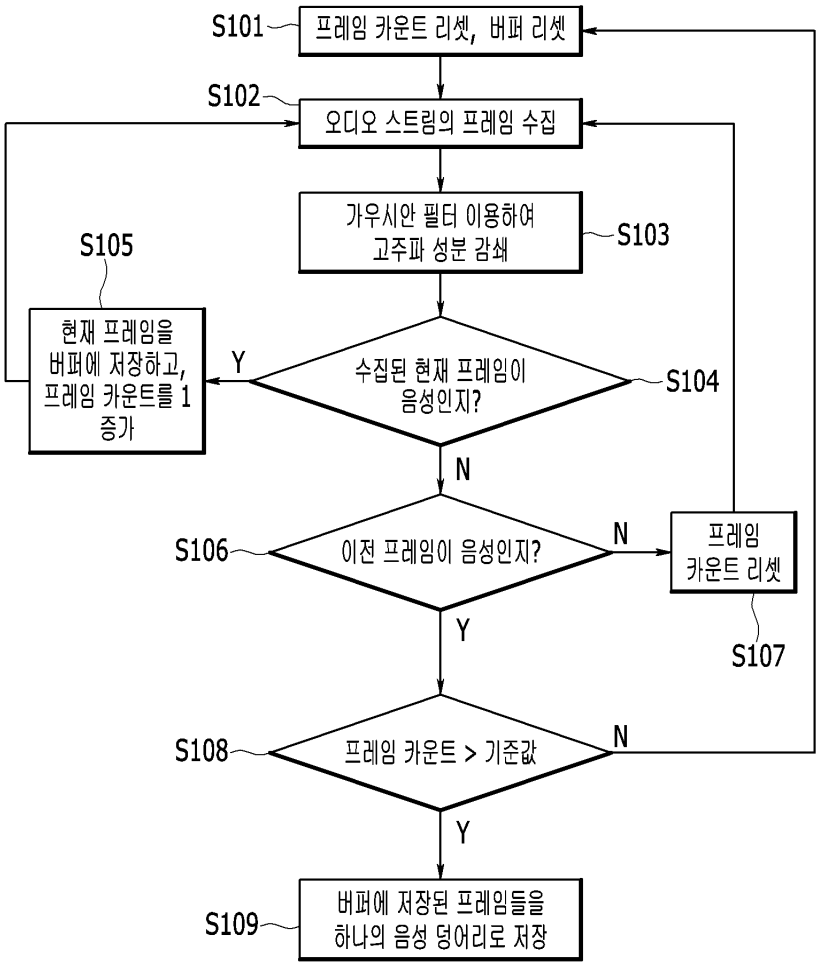
도면2



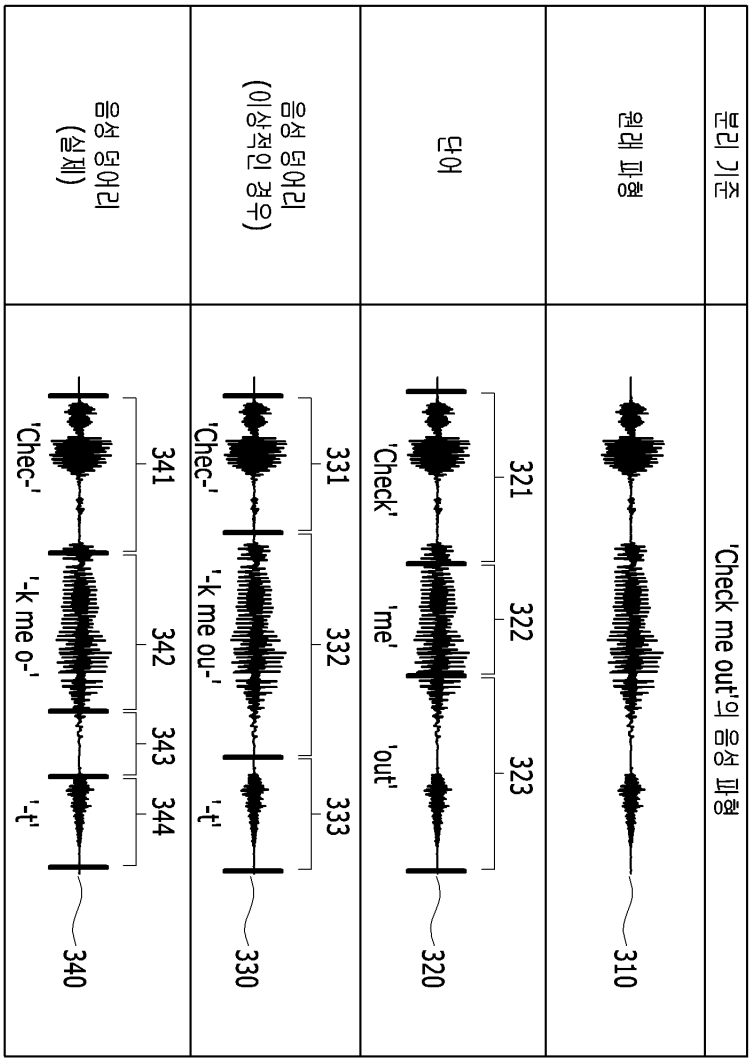
도면3


오디오 스트림
음성 덩어리 분할부
210
음성 덩어리
200
특징 추출부
220
모델 학습부
230
언어 식별 모델
231
언어 식별부
240
식별된 언어 & 확률

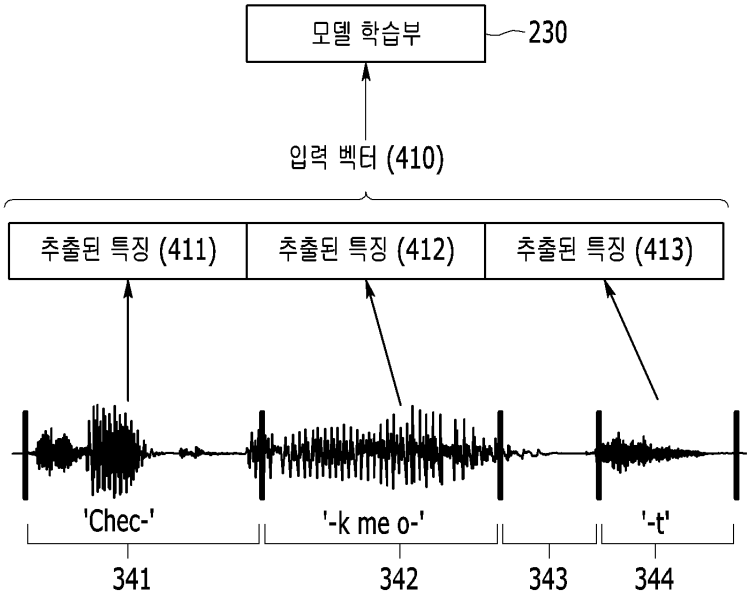
도면4


S101
프레임 카운트 리셋, 버퍼 리셋
S102
오디오 스트림의 프레임 수집
가우시안 필터 이용하여 고주파 성분 감쇄
S103
S105
현재 프레임을 버퍼에 저장하고, 프레임 카운트를 1 증가
Y
수집된 현재 프레임이 음성인지?
S104
N
S106
이전 프레임이 음성인지?
N
프레임 카운트 리셋
S107
Y
S108
프레임 카운트 > 기준값
N
Y
S109
버퍼에 저장된 프레임들을 하나의 음성 덩어리로 저장

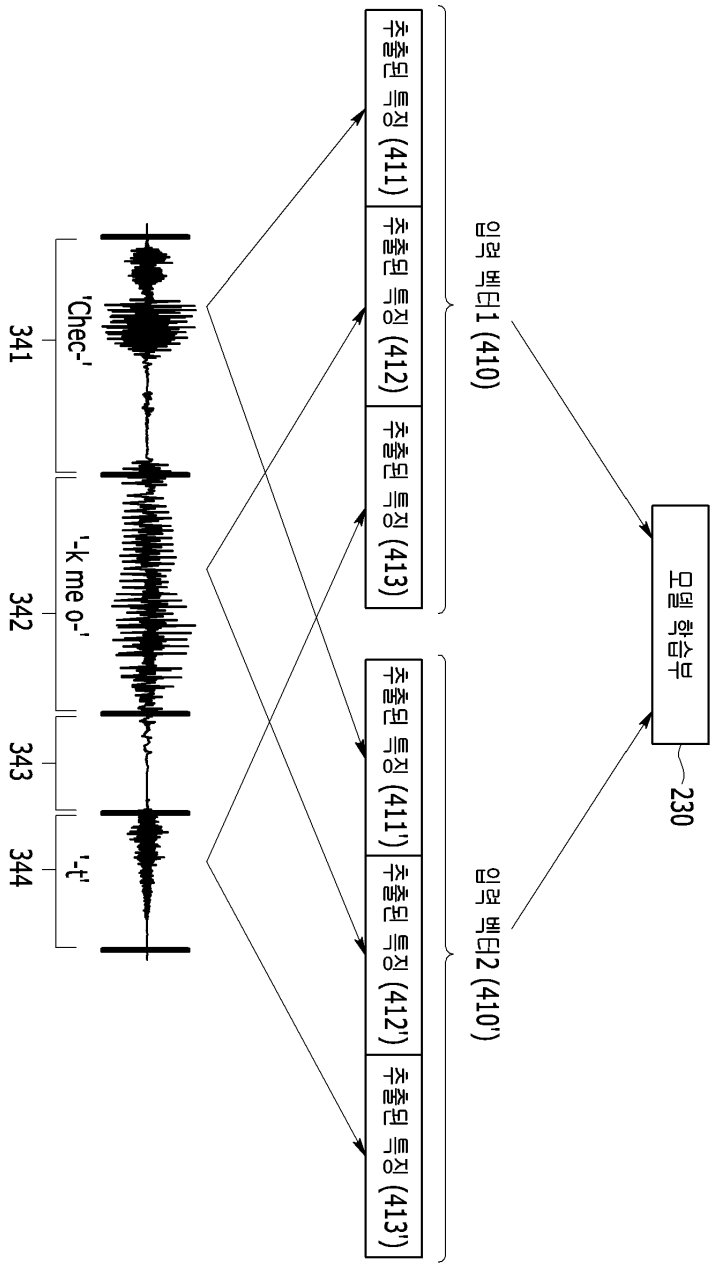
도면5


분리 기준
'Check me out'의 음성 파형
원래 파형
310
단어
320
321
322
323
'Check'
'me'
'out'
음성 덩어리
(이상적인 경우)
330
331
332
333
'Chec-'
'-k me ou-'
'-t'
음성 덩어리
(실제)
340
341
342
343
344
'Chec-'
'-k me o-'
'-t'

도면6


모델 학습부
230
입력 벡터 (410)
추출된 특징 (411)
추출된 특징 (412)
추출된 특징 (413)
'Chec-'
'-k me o-'
'-t'
341
342
343
344

도면7


341
342
343
344
'Chec-'
'-k me o-'
'-t'
추출된 특징 (411)
추출된 특징 (412)
추출된 특징 (413)
입력 벡터1 (410)
추출된 특징 (411')
추출된 특징 (412')
추출된 특징 (413')
입력 벡터2 (410')
모델 학습부
230

도면8