



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0028053
(43) 공개일자 2025년02월28일

(51) 국제특허분류(Int. Cl.)

G10L 25/78 (2013.01) G10L 15/02 (2006.01)

G10L 15/10 (2006.01) G10L 25/93 (2013.01)

(52) CPC특허분류

G10L 25/78 (2013.01)

G10L 15/02 (2013.01)

(21) 출원번호 10-2023-0109214

(22) 출원일자 2023년08월21일

심사청구일자 2023년08월21일

(71) 출원인

계명대학교 산학협력단

대구광역시 달서구 달구벌대로 1095, 계명대학교 산학협력관 201호(신당동)

(72) 발명자

정용주

대구광역시 수성구 범어로20길 60-2 (범어동)

(74) 대리인

특허법인아주

전체 청구항 수 : 총 19 항

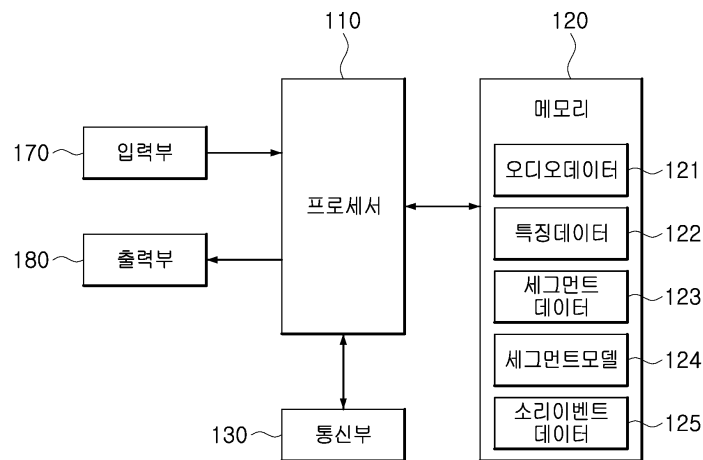
(54) 발명의 명칭 분절모델 기반 소리 이벤트 검출장치 및 그 방법

(57) 요약

본 발명은 분절모델 기반 소리 이벤트 검출장치 및 그 방법에 관한 것으로, 오디오데이터 및 분절모델에 대한 데이터를 저장하는 메모리 및 상기 오디오데이터로부터 길이가 상이한 세그먼트에 대한 데이터를 추출하고, 상기 세그먼트에 대한 유사도 스코어를 산출하며, 상기 분절모델을 기반으로 상기 유사도 스코어를 이용하여 상기 세그먼트에 대한 소리 이벤트를 검출하는 프로세서를 포함하여, 분절모델을 기반으로 오디오데이터를 소리 단위로 분석하여 오디오데이터에 포함된 복수의 소리 또는 묵음 구간을 효과적으로 구분할 수 있고, 불완전 레이블 오류가 포함된 오디오데이터로부터 소리 이벤트를 검출하여 검출 성능을 향상시킬 수 있다.

대표도 - 도1

100



(52) CPC특허분류

G10L 15/10 (2013.01)

G10L 25/93 (2013.01)

명세서

청구범위

청구항 1

오디오데이터 및 분절모델에 대한 데이터를 저장하는 메모리; 및

상기 오디오데이터로부터 길이가 상이한 세그먼트에 대한 데이터를 추출하고, 상기 세그먼트에 대한 유사도 스코어를 산출하며, 상기 분절모델을 기반으로 상기 유사도 스코어를 이용하여 상기 세그먼트에 대한 소리 이벤트를 검출하는 프로세서;를 포함하는 분절모델 기반 소리 이벤트 검출장치.

청구항 2

제 1 항에 있어서,

상기 프로세서는 상기 오디오데이터로부터 추출되는 특징을 인코딩하여, 시간에 대한 상관관계를 모델링하고, 시간 구간의 길이를 단축하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 3

제 2 항에 있어서,

상기 프로세서는, 인코딩 결과로부터 각 세그먼트에 대한 임베딩 벡터를 상기 세그먼트에 대한 데이터로써 획득하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 4

제 3 항에 있어서,

상기 프로세서는 상기 임베딩 벡터를 이용하여, 시작 시점이 t 이고, 세그먼트의 길이가 s 이며 소리의 종류가 y 인 상기 세그먼트에 대한 유사도 스코어를 산출하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 5

제 4 항에 있어서,

상기 프로세서는, 상기 임베딩 벡터에 가중치 벡터를 곱하고 바이어스를 가산하여 상기 세그먼트에 대한 상기 유사도 스코어를 산출하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 6

제 1 항에 있어서,

상기 프로세서는 상기 유사도 스코어가 최대값인 상기 세그먼트를 추출하여 상기 세그먼트에 대한 분할 정보 및 소리의 종류를 획득하고, 상기 세그먼트에 대한 소리 이벤트를 검출하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 7

제 1 항에 있어서,

상기 프로세서는 상기 유사도 스코어에 비용함수를 적용하여 학습을 통해 상기 분절모델에 대한 파라미터를 획득하고, 상기 파라미터를 기반으로 상기 분절모델을 생성하여 상기 메모리에 저장하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 8

제 7 항에 있어서,

상기 프로세서는 상기 오디오데이터에 대한 특징 벡터와 레이블 시퀀스를 기반으로 산출되는 상기 비용함수를

적용하여 학습을 수행하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 9

제 8 항에 있어서,

상기 프로세서는, 상기 오디오데이터에 대한 적어도 하나의 세그먼트를 포함하는 경로를 산출하여 상기 비용함수를 적용한 상기 분절모델을 기반으로,

상기 유사도 스코어가 최대값인 경로를 획득하여, 상기 세그먼트에 대한 분할 정보 및 소리의 종류를 획득하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 10

제 1 항에 있어서,

상기 프로세서는, 임의의 소리 레이블(SIL)을 참레이블 중 하나로 설정하고, 참레이블 비교방식을 통해, 상기 오디오데이터에 포함된 묵음구간 및 복수의 소리를 구분하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 11

프로세서가, 오디오데이터로부터 특징을 추출하여 인코딩하는 단계;

상기 프로세서가, 상기 인코딩 결과에 대한 세그먼트 임베딩을 통해 상기 세그먼트에 대한 데이터를 추출하는 단계;

상기 프로세서가, 상기 세그먼트에 대한 유사도 스코어를 산출하는 단계; 및

상기 프로세서가, 분절모델을 기반으로 상기 유사도 스코어를 이용하여 상기 오디오데이터로부터 소리 이벤트를 검출하는 단계; 를 포함하는 분절모델 기반 소리 이벤트 검출방법.

청구항 12

제 11 항에 있어서,

상기 인코딩하는 단계는,

상기 오디오데이터로부터 추출되는 특징을 인코딩하여, 시간에 대한 상관관계를 모델링하는 단계; 및

상기 특징에 대한 시간 구간의 길이를 단축하는 단계;를 포함하는 분절모델 기반 소리 이벤트 검출방법.

청구항 13

제 11 항에 있어서,

상기 세그먼트에 대한 데이터를 추출하는 단계에서,

상기 프로세서는 인코딩 결과로부터 상기 세그먼트에 대한 임베딩 벡터를 상기 세그먼트에 대한 데이터로써 획득하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출방법.

청구항 14

제 13 항에 있어서,

상기 유사도 스코어를 산출하는 단계에서,

상기 프로세서는 상기 임베딩 벡터를 이용하여, 시작 시점이 t 이고, 세그먼트의 길이가 s 이며 소리의 종류가 y 인 상기 세그먼트에 대한 상기 유사도 스코어를 산출하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출방법.

청구항 15

제 14 항에 있어서,

상기 유사도 스코어를 산출하는 단계에서,

상기 프로세서는 상기 임베딩 벡터에 가중치 벡터를 곱하고 바이어스를 가산하여 상기 세그먼트에 대한 상기 유사도 스코어를 산출하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출방법.

청구항 16

제 11 항에 있어서,

상기 소리 이벤트를 검출하는 단계에서,

상기 프로세서는 상기 유사도 스코어가 최대값인 상기 세그먼트를 추출하여 상기 세그먼트에 대한 분할 정보 및 소리의 종류를 획득하고, 상기 세그먼트에 대한 소리 이벤트를 검출하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출방법.

청구항 17

제 11 항에 있어서,

상기 소리 이벤트를 검출하는 단계에서,

상기 프로세서는, 임의의 소리 레이블(SIL)을 참레이블 중 하나로 설정하고, 참레이블 비교방식을 통해, 상기 오디오데이터에 포함된 묵음구간 및 복수의 소리를 구분하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출장치.

청구항 18

제 11 항에 있어서,

상기 유사도 스코어를 산출하는 단계 후,

상기 프로세서가 상기 유사도 스코어에 비용함수를 적용하여 학습을 통해 상기 분절모델에 대한 파라미터를 획득하는 단계; 및

상기 파라미터를 기반으로 상기 분절모델을 생성하는 단계;를 더 포함하는 분절모델 기반 소리 이벤트 검출방법.

청구항 19

제 18 항에 있어서,

상기 파라미터를 획득하는 단계에서,

상기 프로세서는, 상기 오디오데이터에 대한 특징 벡터와 레이블 시퀀스를 기반으로 산출되는 상기 비용함수를 적용하여 학습을 수행하는 것을 특징으로 하는 분절모델 기반 소리 이벤트 검출방법.

발명의 설명

기술 분야

[0001] 본 발명은 길이가 상이한 세그먼트 단위로 오디오데이터를 분석하여 소리 이벤트를 검출하는 분절모델 기반 소리 이벤트 검출장치 및 그 방법에 관한 것이다.

배경 기술

[0002] 소리 이벤트 검출(Sound event detection)은 특정한 소리의 발생 여부 및 해당 소리의 발생 시점을 탐지하는 기법이다.

[0003] 전통적으로는 GMM(Gaussian Mixture Model) 또는 SVM(Support Vector Machine)을 기반으로 한 방식들이 많이 사용되어 왔으나, 최근에는 기존의 머신러닝 기반의 방식보다 뛰어난 성능을 보여주는 딥뉴럴네트워크(Deep neural network) 기반 방식들이 많이 사용되는 추세이다.

- [0004] 소리 이벤트 검출 기법은, 소리를 이용한 감시, 도심지 잡음 해석, 멀티미디어 콘텐츠로부터의 정보 탐색, 헬스케어 모니터링 및 새소리 탐지 등 다양한 분야에서 활용되고 있다.
- [0005] 소리 이벤트 검출 기법은 다양한 방식들이 제안되었고, 최근에는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network)을 결합한 방식인 CRNN(Convolutional Recurrent Neural Network)이 대표적으로 사용되고 있다.
- [0006] 그러나 CRNN을 이용한 소리 이벤트 검출 기법은, 신뢰할 만한 학습 데이터를 확보가 어려워 검출 성능이 저하되는 문제가 있다.
- [0007] 특히, 웹 기반 오디오데이터는 대용량의 오디오데이터의 쉽게 확보할 수 있으나, 많은 양의 레이블 오류를 포함하는 문제가 있다.
- [0008] 대표적인 레이블 오류는 부정확(Incorrect) 레이블 오류 및 불완전(Incomplete) 레이블 오류이다.
- [0009] 부정확 레이블 오류는 전사자(Annotator)의 착오나 소리들 간의 의미론적인 애매함으로 인한 다수의 레이블 오류가 존재하는 경우이다.
- [0010] 불완전 레이블 오류는, 하나의 오디오 파일에 다수의 소리가 존재하거나 긴 묵음 구간이 소리와 함께 존재하는 경우, 소리들 간의 정확한 경계가 알려지지 않은 경우이다.
- [0011] 부정확 레이블 오류에 대해서는 그 대처 방안에 관하여 많은 연구가 진행되고 있으나, 불완전 레이블 오류에 관한 연구는 거의 진행되지 않은 실정이다.
- [0012] 그에 따라 불완전 레이블 오류를 고려하여 효과적으로 소리 이벤트를 검출하는 방안이 필요하다.
- [0013] 관련 기술로 대한민국 등록특허공보 제10-2374144호, "인공지능 기반의 이상음원 인식 장치, 그 방법 및 이를 이용한 관제시스템"이 있다.

발명의 내용

해결하려는 과제

- [0014] 본 발명은 상기와 같은 필요성에 의해 창출된 것으로서, 분절모델을 기반으로 불완전 레이블 오류를 가진 오디오데이터로부터 소리 이벤트를 검출하는, 분절모델 기반 소리 이벤트 검출장치 및 그 방법을 제공하는 데 그 목적이 있다.

과제의 해결 수단

- [0015] 상기한 목적을 달성하기 위하여 본 발명에 따른 분절모델 기반 소리 이벤트 검출장치는, 오디오데이터 및 분절모델에 대한 데이터를 저장하는 메모리; 및 상기 오디오데이터로부터 길이가 상이한 세그먼트에 대한 데이터를 추출하고, 상기 세그먼트에 대한 유사도 스코어를 산출하며, 상기 분절모델을 기반으로 상기 유사도 스코어를 이용하여 상기 세그먼트에 대한 소리 이벤트를 검출하는 프로세서;를 포함한다.
- [0016] 상기 프로세서는 상기 오디오데이터로부터 추출되는 특징을 인코딩하여, 시간에 대한 상관관계를 모델링하고, 시간 구간의 길이를 단축하는 것을 특징으로 한다.
- [0017] 상기 프로세서는, 인코딩 결과로부터 각 세그먼트에 대한 임베딩 벡터를 상기 세그먼트에 대한 데이터로써 획득하는 것을 특징으로 한다.
- [0018] 상기 프로세서는 상기 임베딩 벡터를 이용하여, 시작 시점이 t 이고, 세그먼트의 길이가 s 이며 소리의 종류가 y 인 상기 세그먼트에 대한 유사도 스코어를 산출하는 것을 특징으로 한다.
- [0019] 상기 프로세서는, 상기 임베딩 벡터에 가중치 벡터를 곱하고 바이어스를 가산하여 상기 세그먼트에 대한 상기 유사도 스코어를 산출하는 것을 특징으로 한다.
- [0020] 상기 프로세서는 상기 유사도 스코어가 최대값인 상기 세그먼트를 추출하여 상기 세그먼트에 대한 분할 정보 및 소리의 종류를 획득하고, 상기 세그먼트에 대한 소리 이벤트를 검출하는 것을 특징으로 한다.
- [0021] 상기 프로세서는 상기 유사도 스코어에 비용함수를 적용하여 학습을 통해 상기 분절모델에 대한 파라미터를 획득하고, 상기 파라미터를 기반으로 상기 분절모델을 생성하여 상기 메모리에 저장하는 것을 특징으로 한다.

- [0022] 상기 프로세서는 상기 오디오데이터에 대한 특징 벡터와 레이블 시퀀스를 기반으로 산출되는 상기 비용함수를 적용하여 학습을 수행하는 것을 특징으로 한다.
- [0023] 상기 프로세서는, 상기 오디오데이터에 대한 적어도 하나의 세그먼트를 포함하는 경로를 산출하여 상기 비용함수를 적용한 상기 분절모델을 기반으로, 상기 유사도 스코어가 최대값인 경로를 획득하여, 상기 세그먼트에 대한 분할 정보 및 소리의 종류를 획득하는 것을 특징으로 한다.
- [0024] 상기 프로세서는, 임의의 소리 레이블(SIL)을 참레이블 중 하나로 설정하고, 참레이블 비교방식을 통해, 상기 오디오데이터에 포함된 묵음구간 및 복수의 소리를 구분하는 것을 특징으로 한다.
- [0025] 본 발명의 분절모델 기반 소리 이벤트 검출방법은, 프로세서가, 오디오데이터로부터 특징을 추출하여 인코딩하는 단계; 상기 프로세서가, 상기 인코딩 결과에 대한 세그먼트 임베딩을 통해 상기 세그먼트에 대한 데이터를 추출하는 단계; 상기 프로세서가, 상기 세그먼트에 대한 유사도 스코어를 산출하는 단계; 및 상기 프로세서가, 분절모델을 기반으로 상기 유사도 스코어를 이용하여 상기 오디오데이터로부터 소리 이벤트를 검출하는 단계;를 포함한다.
- [0026] 상기 인코딩하는 단계는, 상기 오디오데이터로부터 추출되는 특징을 인코딩하여, 시간에 대한 상관관계를 모델링하는 단계; 및 상기 특징에 대한 시간 구간의 길이를 단축하는 단계;를 포함한다.
- [0027] 상기 세그먼트에 대한 데이터를 추출하는 단계에서, 상기 프로세서는 인코딩 결과로부터 상기 세그먼트에 대한 임베딩 벡터를 상기 세그먼트에 대한 데이터로써 획득하는 것을 특징으로 한다.
- [0028] 상기 유사도 스코어를 산출하는 단계에서, 상기 프로세서는 상기 임베딩 벡터를 이용하여, 시작 시점이 t 이고, 세그먼트의 길이가 s 이며 소리의 종류가 y 인 상기 세그먼트에 대한 상기 유사도 스코어를 산출하는 것을 특징으로 한다.
- [0029] 상기 유사도 스코어를 산출하는 단계에서, 상기 프로세서는 상기 임베딩 벡터에 가중치 벡터를 곱하고 바이어스를 가산하여 상기 세그먼트에 대한 상기 유사도 스코어를 산출하는 것을 특징으로 한다.
- [0030] 상기 소리 이벤트를 검출하는 단계에서, 상기 프로세서는 상기 유사도 스코어가 최대값인 상기 세그먼트를 추출하여 상기 세그먼트에 대한 분할 정보 및 소리의 종류를 획득하고, 상기 세그먼트에 대한 소리 이벤트를 검출하는 것을 특징으로 한다.
- [0031] 상기 소리 이벤트를 검출하는 단계에서, 상기 프로세서는, 임의의 소리 레이블(SIL)을 참레이블 중 하나로 설정하고, 참레이블 비교방식을 통해, 상기 오디오데이터에 포함된 묵음구간 및 복수의 소리를 구분하는 것을 특징으로 한다.
- [0032] 상기 유사도 스코어를 산출하는 단계 후, 상기 프로세서가 상기 유사도 스코어에 비용함수를 적용하여 학습을 통해 상기 분절모델에 대한 파라미터를 획득하는 단계; 및 상기 파라미터를 기반으로 상기 분절모델을 생성하는 단계;를 더 포함한다.
- [0033] 상기 파라미터를 획득하는 단계에서, 상기 프로세서는, 상기 오디오데이터에 대한 특징 벡터와 레이블 시퀀스를 기반으로 산출되는 상기 비용함수를 적용하여 학습을 수행하는 것을 특징으로 한다.

발명의 효과

- [0034] 일 측면에 따르면, 본 발명의 분절모델 기반 소리 이벤트 검출장치 및 그 방법은, 불완전 레이블 오류가 포함된 오디오데이터를 분절모델을 기반으로 분석하여 소리 이벤트를 정확하게 검출하고, 검출 성능을 향상시킬 수 있다.
- [0035] 발명의 일 측면에 따른 분절모델 기반 소리 이벤트 검출장치 및 그 방법은, 세그먼트 기반으로 오디오데이터를 분석하여 연산량을 감소시킬 수 있고, 그에 따른 효율성을 향상시킬 수 있다.
- [0036] 본 발명의 일 측면에 따른 분절모델 기반 소리 이벤트 검출장치 및 그 방법은, 분절모델을 기반으로 오디오데이터를 분석하여, 구간 정보를 제공하지 않으면서 오디오데이터에 포함된 복수의 소리 또는 묵음 구간을 효과적으로 구분할 수 있다.
- [0037]

도면의 간단한 설명

- [0038] 도 1 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 구성이 도시된 블록 구성도이다.
- 도 2 는 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 오디오 특징 추출에 따른 데이터 흐름이 도시된 흐름도다.
- 도 3 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 분절모델 생성에 관한 데이터 흐름이 도시된 흐름도이다.
- 도 4 는 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 오디오데이터의 소리 이벤트 검출을 위한 신호구성이 도시된 도이다.
- 도 5 는 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 불완전 레이블 오류 검출을 설명하는데 참조되는 도이다.
- 도 6 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 분절모델 생성방법이 도시된 순서도이다.
- 도 7 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 분절모델 기반 소리 이벤트 검출방법이 도시된 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0039] 이하, 첨부된 도면들을 참조하여 본 발명을 설명하도록 한다.
- [0040] 이 과정에서 도면에 도시된 선들의 두께나 구성요소의 크기 등은 설명의 명료성과 편의상 과장되게 도시되어 있을 수 있다. 또한, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로써 이는 사용자, 운용자의 의도 또는 관례에 따라 달라질 수 있다. 그러므로 이러한 용어들에 대한 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [0041] 도 1 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 구성이 도시된 블록 구성도이다.
- [0042] 도 1에 도시된 바와 같이, 소리 이벤트 검출장치(이하, 검출장치)(100)는 통신부(130), 메모리(120), 입력부(170), 출력부(180) 및 프로세서(110)를 포함한다.
- [0043] 메모리(120)는 입력부(170)를 통해 입력되어 저장되는 오디오데이터(121), 오디오데이터(121)로부터 추출되는 특징데이터(122), 소리 단위의 세그먼트데이터(123), 세그먼트데이터에 대한 학습모델인 세그먼트모델(124) 및 세그먼트모델(124)을 기반으로 오디오데이터(121)로부터 추출되는 소리이벤트데이터(125)를 저장할 수 있다.
- [0044] 메모리(120)는 오디오데이터(121)를 신호처리하고 분석하기 위한 데이터를 더 포함할 수 있다.
- [0045] 또한, 메모리(120)는 인코딩 알고리즘, 디코딩 알고리즘, 학습 알고리즘, 세그먼트 분석 알고리즘, 스코어링 알고리즘, 및 소리 이벤트 검출 알고리즘 중 적어도 하나에 대한 데이터를 저장할 수 있다.
- [0046] 예를 들어, 메모리(120)는 램(RAM, Random Access Memory), 롬(ROM), EEPROM(Electrically Erasable Programmable ROM) 등의 비휘발성 메모리, 플래시 메모리 등의 저장수단을 포함할 수 있다.
- [0047] 통신부(130)는 유선 또는 무선의 통신모듈을 포함하여 검출장치(100)와 연결되는 장치와 통신할 수 있다. 통신부(130)는 네트워크를 통해 연결되는 단말(미도시) 및 서버(미도시)와 통신할 수 있다.
- [0048] 예를 들어 통신부(130)는 이더넷(Ethernet), 와이파이(WIFI), 블루투스(BLUETOOTH) 등의 근거리 통신, 이동 통신, 및 시리얼 통신 중 적어도 하나를 포함할 수 있다.
- [0049] 입력부(170)는 버튼, 스위치, 및 터치패드 중 적어도 하나의 입력수단을 포함하여 사용자 명령을 입력한다. 입력부(170)는 입력수단을 통해 입력되는 오디오데이터(121)에 대한 신호처리 및 소리 이벤트 검출과 관련하여 사용자 명령을 입력할 수 있다.
- [0050] 또한, 입력부(170)는 오디오데이터(121)를 입력받는 입력인터페이스를 포함할 수 있다. 예를 들어 입력부(170)는 마이크(마이크) 또는 오디오 입력단자를 포함할 수 있다.
- [0051] 출력부(180)는 스피커, 동작램프 및 디스플레이 중 적어도 하나를 포함할 수 있다.
- [0052] 출력부(180)는 입력된 오디오데이터(121)의 파형을 디스플레이를 통해 출력하고, 스피커를 통해 소리를 출력할

수 있다.

- [0053] 출력부(180)는 오디오데이터(121)의 신호처리 및 소리 이벤트 검출에 대한 진행상태와 그 결과를 디스플레이를 통해 출력할 수 있다. 출력부(180)는 오디오데이터(121)에 대한 소리 이벤트 검출에 관한 메뉴화면을 디스플레이를 통해 출력할 수 있다.
- [0054] 출력부(180)는 진행상태 및 결과에 대응하여 스피커를 통해 소정의 효과음 또는 경고음을 출력할 수 있다. 또한, 출력부(180)는 진행상태를 동작램프를 통해 출력할 수 있다.
- [0055] 프로세서(110)는 입력되는 오디오데이터(121)를 메모리(120)에 저장하고, 메모리(120)에 저장된 데이터를 호출하여 신호처리를 수행하고 오디오데이터(121)에 대한 특징을 추출하여 메모리(120)에 특징데이터(122)로 저장할 수 있다.
- [0056] 프로세서(110)는 추출된 특징을 기반으로 오디오데이터(121)의 세그먼트를 구분한다.
- [0057] 프로세서(110)는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network) 및 LSTM(Long Short Term Memory)을 기반으로 특징데이터(122)를 인코딩(Encoding) 할 수 있다.
- [0058] 프로세서(110)는 세그먼트를 구분하고 딥러닝 기반 학습을 통해 세그먼트모델(분절모델)을 생성할 수 있다. 프로세서(110)는 세그먼트데이터(123) 및 세그먼트모델(124)을 메모리(120)에 저장한다.
- [0059] 프로세서(110)는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network)을 결합한 방식인 CRNN(Convolutional Recurrent Neural Network)을 이용하여 세그먼트데이터(123)를 추출할 수 있다.
- [0060] 또한, 프로세서(110)는 기 생성된 세그먼트모델(분절모델)을 기반으로 오디오데이터(121)를 분석하여 세그먼트를 구분하고, 이를 기반으로 소리 이벤트를 검출할 수 있다. 프로세서(110)는 검출된 소리 이벤트 데이터를 메모리(120)에 저장한다.
- [0061] 프로세서(110)는 오디오데이터(121)로부터 추출된 특징에 대하여 프레임 단위가 아닌 소리 단위, 즉 세그먼트 단위로 데이터를 처리한다. 이때 세그먼트는 소리에 따라 길이가 상이하게 설정된다.
- [0062] 프로세서(110)는 소리의 정확한 길이를 알 수 없으므로, 오디오데이터(121)로부터 추출되는 특징으로부터 임베딩되는 세그먼트에 대하여 유사도 스코어링을 수행한다. 프로세서(110)는 이러한 과정을 반복하여 학습함으로써, 세그먼트모델(124)을 생성할 수 있다.
- [0063] 또한, 프로세서(110)는 생성된 세그먼트모델(124)을 기반으로 오디오데이터(121)를 세그먼트 단위로 구분하고, 각 세그먼트의 길이를 최적화할 수 있다. 그에 따라 프로세서(110)는 오디오데이터(121)에 포함된 소리 이벤트를 검출할 수 있다.
- [0064] 프로세서(110)는 하나의 오디오데이터(121)에 복수의 소리가 포함된 경우 세그먼트를 추출하고, 세그먼트에 대하여 임의의 소리 레이블(SIL)을 적용하여 참레이블 비교 방식을 통해 복수의 소리를 각각 구분하여, 소리 이벤트로 검출할 수 있다.
- [0065] 또한, 프로세서(110)는 오디오데이터(121)에 묵음구간이 포함된 경우 세그먼트에 대한 스코어링을 통해 세그먼트를 추출하고, 임의의 소리 레이블(SIL)을 적용하여 참레이블 비교 방식을 통해 묵음구간을 검출하고, 묵음구간을 배제한 나머지 구간에서 소리 이벤트를 검출할 수 있다.
- [0066] 프로세서(110)는 오디오데이터(121)에 포함된 소리의 세그먼트 정보를 알 수 없는 상태에서 세그먼트모델(124)을 기반으로 세그먼트 단위로 데이터를 처리하여 구간 정보 없이 소리 이벤트를 검출할 수 있다.
- [0067] 도 2 는 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 오디오 특징 추출에 따른 데이터 흐름이 도시된 흐름도다.
- [0068] 도 2에 도시된 바와 같이, 검출장치(100)는 오디오데이터(121)로 소리 이벤트를 검출하기 위해 특징을 추출한 후 세그먼트모델의 학습 및 디코딩을 수행할 수 있다.
- [0069] 검출장치(100)는 다음과 같이 오디오데이터(121)로부터 특징을 추출할 수 있다. 프로세서(110)는 추출되는 특징을 특징데이터(122)로 메모리(120)에 저장한다.
- [0070] 입력부(170)를 통해 오디오데이터(121)가 입력되면(S10), 프로세서(110)는 해밍윈도우(HAMMING WINDOW)를 사용하여, 시간에 따른 오디오데이터(121)를 분석한다(S20).

- [0071] 프로세서(110)는 구간을 설정하여 윈도우 별로 신호를 분석할 수 있다. 오디오데이터(121)는 16KHz의 샘플링 레이트가 사용될 수 있다.
- [0072] 프로세서(110)는 약 41.5ms의 프레임 간격마다 STFT(Short-Time Fourier Transform)을 연산한다(S30).
- [0073] 프로세서(110)는 STFT 값으로부터 멜 필터링(Mel-Filtering)을 수행하고(S40), 그 결과를 로그 변환하여(S50), 로그-멜 필터뱅크(Log-Mel Filterbank)를 획득한다(S60).
- [0074] 프로세서(110)는 64 band의 멜-필터뱅크를 전체 0에서 8,000Hz의 주파수 구간에서 산출한 후, 이를 로그 변환함으로써 64차원의 로그-멜 필터뱅크를 프레임마다 획득할 수 있다.
- [0075] 프레임 간격 및 로그-멜-필터뱅크 값의 차원에 대한 설정은 DCASE 2018 오디오데이터(121)에 대한 추천값(J. Y. Kwak and Y. J. Chung, "Sound Event Detection Based on CRNN Using Derivative Features", Journal of KIIT, Vol. 18, No. 6, pp. 89-96, Jun. 2020. 18.6.89. 참조)을 적용하였으나, 이는 일 예일 뿐 설정에 따라 변경될 수 있다.
- [0076] 프로세서(110)는 로그-멜 필터뱅크에서 학습 데이터 전체의 평균값을 뺀 후, 표준편차로 나누어 정규화함으로써 오디오데이터(121)로부터 특징을 추출할 수 있다.
- [0077] 도 3 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 분절모델 생성에 관한 데이터 흐름이 도시된 흐름도이다.
- [0078] 도 3에 도시된 바와 같이, 검출장치(100)는 오디오데이터(121)로부터 추출된 특징을 기반으로 세그먼트모델을 생성하고, 디코딩을 통해 소리 이벤트를 추출할 수 있다.
- [0079] 검출장치(100)는 오디오데이터(121)로부터 특징이 추출되면(Feature extraction)(S110), 인코딩(Encoding)을 수행하고(S120), 인코딩된 데이터로부터 세그먼트 임베딩(Embedding)(S130)을 수행한다. 또한, 검출장치(100)는 세그먼트에 대한 유사도 스코어링(Scoring)(S140)을 수행할 수 있다.
- [0080] 검출장치(100)는 손실함수 및 역전파를 획득하여(S150), 학습을 통해 세그먼트모델(분절모델)(124)을 생성할 수 있다.
- [0081] 또한, 검출장치(100)는 세그먼트모델(분절모델)(124)을 기반으로 세그먼트 스코어링 결과로부터 세그먼트를 최적화할 수 있다(S160).
- [0082] 특징 추출(Feature extraction)(S110)이 완료되면, 프로세서(110)는 특징 추출과정에서 획득된 로그-멜 필터뱅크의 출력 $x=\{x_1, x_2, x_3, \dots, x_T\}$, $x_t \in R^B$ 을 인코딩(Encoding)한다(S120).
- [0083] 프로세서(110)는 약 64개의 은닉 뉴런을 가진 bi-directional LSTM(Long Short-Term Memory)과 5x5의 커널 사이즈를 가진 CNN으로 구성하여, 특징을 인코딩할 수 있다.
- [0084] 프로세서(110)는 LSTM을 통해 오디오데이터(121)에 포함된 신호들의 긴 시간의 상관관계를 모델링하고, CNN의 풀링(pooling)을 통하여 시간 구간 길이를 축소함으로써 불필요한 정보를 제한하고 계산량을 감소시킬 수 있다.
- [0085] 프로세서(110)는 추출된 특징인 로그-멜 필터뱅크의 출력 x 를 인코딩하여 새로운 특징 $h=Enc(x)$, $h=\{h_1, h_2, h_3, \dots, h_T\}$, $h \in R^F$ 을 생성할 수 있다.
- [0086] 이때, B는 로그-멜 필터뱅크의 벡터이고, F는 인코딩 벡터의 차원이고, T는 오디오데이터(121)의 전체 프레임 길이이며, $\hat{T}$ 는 인코딩 벡터의 전체 길이이다.
- [0087] 프로세서(110)는 인코딩 결과를 기반으로, 세그먼트 임베딩(Embedding)(S130)을 수행한다.
- [0088] 프로세서(110)는 인코딩 결과인 h 로부터 각 세그먼트에 적합한 임베딩 벡터를 추출한다.
- [0089] 예를 들어, 프로세서(110)는 시간 구간 $[t, t+s]$ 에 해당하는 임의의 세그먼트 $h_{t:t+s} \in R^{s \times F}$ 와 관련하여 임베딩 벡터 $I(h_{t:t+s}) \in R^D$ 를 획득할 수 있다.
- [0090] 프로세서(110)는 다음 수학적 식 1과 같이 ReLU 활성화함수를 가지는 FNN(Feed-forward Neural Network)을 이용하여

임베딩 벡터를 획득할 수 있다.

수학식 1

$$I(h_{t,t+s})=W_1P(h_{t,t+s})+b_1$$

여기서 $P(\cdot)$ 는 주어진 세그먼트에 대한 풀링(pooling)을 의미하며, 아래 수학식 2와 같이 각 세그먼트의 처음과 끝의 인코딩 벡터의 단순한 결합으로 사용될 수 있다.

여기서 W_1 은 FNN의 가중치 벡터이고, b_1 은 FNN의 바이어스(Bias)를 나타낸다.

수학식 2

$$P(h_{t,t+s})=[h_t;h_{t+s}]$$

프로세서(110)는 수학식 2와 같이, 각 세그먼트에 대한 임베딩 벡터를 획득할 수 있다.

프로세서(110)는 획득된 임베딩 벡터를 이용하여 시작 시점이 t 이고 그 길이가 s 이며, 소리의 종류가 y 인 세그먼트에 대한 유사도 스코어(Score)인 $w_{t,s,y}$ 를 수학식 3과 같이 내적(Dot product)을 이용하여 산출한다.

수학식 3

$$w_{t,s,y}=W_2I(h_{t,t+s})+b_2$$

$$1 \leq t < \hat{T}, \quad 1 \leq s < S, \quad 1 \leq y < Y$$

여기서, S 는 하나의 세그먼트의 가능한 최대 길이이고, Y 는 분류하고자 하는 소리 종류의 전체 개수이며, W_2 는 가중치 벡터이고, b_2 는 바이어스를 나타낸다.

세그먼트의 길이 s 는 1 이상이고 세그먼트의 최대 길이보다 작은 값으로 설정된다. 시작 시점 t 는 1 이상이고 오디오데이터(121)의 전체 프레임 길이 T 보다 작게 설정된다. 또한, 소리의 종류 y 는 1 이상이고, 분류하고자 하는 소리의 전체 개수보다 작은 값이다.

세그먼트모델 생성을 위한 학습 단계에서, 프로세서(110)는 학습을 위한 비용함수로써, Marginal Log Loss를 사용할 수 있다.

프로세서(110)는 비용함수를 이용하여, 오디오 분류기의 성능 저하를 발생시키는 불완전(Incomplete) 레이블 오류를 처리할 수 있다.

학습데이터로써 입력되는 특징벡터 $x=\{x_1,x_2,x_3,\dots,x_T\}$ 와 레이블 시퀀스(Sequence) $Y=\{y_1,y_2,y_3,\dots,y_L\}$ 이 주어질 경우, 비용함수는 다음 수학식 4와 같이 정의될 수 있다.

수학식 4

$$\mathcal{L}(Y,x)=-\log p(Y|x)=-\log \sum_z p(Y,z|x)$$

여기서, z 는 입력 x 가 가질 수 있는 모든 가능한 시간 분할(Segmentation)을 의미한다.

프로세서(110)는 수학식 4로부터 다음 수학식 5를 산출할 수 있다.

수학식 5

$$p(Y, z | x) = p(\pi | x) = \frac{1}{Z(x)} \exp(w(x, \pi))$$

여기서, π 는 $\{Y, z\}$ 로 구성되는 입력오디오 x 에 대한 하나의 경로(Path)를 나타낸다. $w(x, \pi)$ 는 해당 경로에 대한 전체 스코어를 의미한다. 이때 $w(x, \pi)$ 는 n 경로를 구성하는 각 세그먼트에 대하여 수학식 3을 이용하여 스코어를 각각 산출한 후 이를 합산하여 획득할 수 있다.

비용함수 $\mathcal{L}(Y, x)$ 는 수학식 4 및 수학식 5를 기반으로 다음 수학식 6과 같이 정의 될 수 있다.

수학식 6

$$\begin{aligned} \mathcal{L}(Y, x) &= -\log \sum_z p(Y, z | x) \\ &= -\log \sum_z \exp(w(x, (Y, z))) + \log X(x) \\ &= -\log \sum_{\pi} \exp(w(x, \pi)) + \log X(x) \end{aligned}$$

$$L(\bar{\pi}) = Y$$

$$Z(x) = \sum_{\pi} \exp(w(x, \pi))$$

여기서 $L(\bar{\pi}) = Y$ 는 해당 경로 $\bar{\pi}$ 가 레이블 시퀀스 Y 를 가짐을 의미한다.

프로세서(110)는 수학식 6의 비용함수 및 그에 대한 기울기 하강(gradient descent) 값을 Forward-Backward 알 고리즘을 이용하여 연산할 수 있다.

프로세서(110)는 연산 결과를 기반으로 반복적 학습을 수행하여 분절 모델을 구성하는 뉴럴네트워크의 파라미터 를 획득할 수 있다.

디코딩 단계에서, 프로세서(110)는 다음 수학식 7과 같이 최적의 경로를 산출할 수 있다.

수학식 7

$$p^* = \arg \max_{\pi} w(x, \pi)$$

프로세서(110)는 수학식 7을 기반으로 스코어가 가장 큰 최적의 경로 $p^* = \{Y^*, z^*\}$ 을 획득할 수 있다.

획득된 최적의 경로를 통해, 프로세서(110)는 최적의 분할 정보와 그에 해당하는 소리의 종류를 확인할 수 있다. 이때 수학식 7은 비터비(Viterbi) 알고리즘을 이용한 동적프로그래밍을 통하여 효과적으로 연산할 수 있 다.

도 4는 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 오디오데이터(121)의 소리 이벤트 검출을 위한 신 호구성이 도시된 도이다.

도 4에 도시된 바와 같이, 검출장치(100)는 marginal log loss의 비용함수를 적용하여 소리 이벤트를 검출할 수 있다.

이때, 검출장치(100)는 학습데이터의 분할정보는 사용하지 않고, 소리 이벤트 검출 실험을 수행할 수 있다.

이때, 도 4와 같이, 소리 이벤트 검출을 위해 시험에 사용된 오디오데이터(121)는 각 파일의 길이가 10초이고,

분류하고자 하는 소리의 종류는 10개이며 가정에서 흔히 발생하는 소리 이벤트로 구성되어 있다.

- [0121] 오디오데이터(121)는, 각각 학습(Train), 테스트(Test) 및 검증(Validation)에 대한 데이터로 구성될 수 있다. 학습데이터는 2548개의 파일을 포함하고, 테스트데이터는 약 692개의 파일을 포함하며, 검증데이터는 약 1168개의 파일을 포함한다.
- [0122] 검출장치(100)는 이와 같은 오디오데이터(121)를 기반으로 소리 이벤트를 검출할 수 있다.
- [0123] 검출장치(100)의 소리 이벤트 검출 결과에 대한 성능은 F 스코어(F-score)를 이용하여, 이벤트 기반(Event-based) 분석방법을 적용하여 평가할 수 있다. 이벤트 기반 분석방법은 검출장치(100)에서 특정 이벤트가 발생한 경우, 그 결과를 참 레이블 정보(Ground truth)와 비교하는 방법이다.
- [0124] 초기 판단은 TP(True Positive), FP(False Positive), FN(False Negative)의 3가지 형태로 수행할 수 있다.
- [0125] TP는 검출된 소리 이벤트와 참 레이블 정보상의 시작 시간과 종료 시간이 겹치는 경우에 해당한다. TP의 판단시, 시작 시간과 종료 시간에서 각각 200ms 오차가 허용되나, 겹치는 구간이 전체 소리 이벤트 길이의 20%를 넘어야 한다.
- [0126] FP는 TP와 상반되는 개념으로, 검출장치(100)에 의해 소리 이벤트가 발생됨에도 불구하고 참 레이블 정보와 겹치는 구간이 발생하지 않는 경우이다. FN는 참 레이블 정보에는 소리 이벤트가 존재하나 검출장치(100)의 출력이 존재하지 않는 경우이다.
- [0127] F 스코어(F-score)는 위에서 언급된 3가지 초기 판단을 근거로 계산되며 Precision과 Recall의 조화평균 값이다. Precision은 참인 문제에 대해 얼마나 잘 맞추었는지에 대한 값이며, Recall은 참인 문제에 대해 정확히 양성으로 식별한 비율을 말한다.
- [0128] F 스코어(F-score)는 수학적 식 8과 같이 산출할 수 있다.

수학적 식 8

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}$$

$$F = \frac{2 \cdot P \cdot R}{P+R}$$

- [0129]
- [0130] 소리 이벤트 검출 시, 검출장치(100)는 오디오데이터(121)의 학습 및 인식 과정에서 적용되는 세그먼트의 최대 가능 길이는 파일 길이와 동일하게 10초로 설정하여 긴 지속시간을 가진 소리들에 대비하도록 설정할 수 있다.
- [0131] 소리 이벤트 검출 시, Batch size는 16으로 설정하였으며, 모델 훈련은 marginal log loss를 손실함수로 삼아 Adam optimizer를 이용하고, 학습률(Learning rate)은 0.001로 설정할 수 있다.
- [0132] 검출장치(100)는 검증데이터에 대한 F 스코어(F-score)값을 기준으로 early stopping을 적용하여 학습을 수행할 수 있다.
- [0133] 표 1 은 RNN 층수에 따른 F 스코어(F-score)의 성능변화를 나타낸 것이다.

표 1

RNNs의 층수	Test set	Validation set	Train set
1	14.87	9.23	57.07
2	10.49	9.26	42.44
3	13.91	8.83	45.55
4	11.48	7.92	33.65

- [0135] 표 1에 기재된 바와 같이, 제안된 세그먼트모델(분절모델)에서 RNN의 층수를 변화시키면서 검출장치(100)의 성능을 테스트할 수 있다.

- [0136] 검출장치(100)는 RNN의 층수가 증가하는 경우 F 스코어(F-score)값이 전반적으로 하락하는 결과가 도출되었다.
- [0137] 이는 오디오데이터(121)의 양에 비해서 RNN 파라미터의 수가 많다는 것을 의미한다.
- [0138] 검출장치(100)는 RNN의 층수가 1일 때 최고의 성능을 얻을 수 있었다. 소리 이벤트 검출에 대한 성능 평가 결과, 테스트데이터에 대해서는 14.87%, 검증데이터와 학습데이터에 대해서는 각각 9.23%와 57.07%의 F 스코어(F-score)값을 얻을 수 있었다.
- [0139] 다음 표 2는 CRNN 기반의 검출장치와 본 발명의 분절모델 기반의 검출장치의 성능을 비교한 것이다.
- [0140] 표 2는 본 발명의 분절모델을 이용한 검출장치와 CRNN 기반의 검출장치가, 동일한 강전사 레이블의 학습데이터를 사용하고, DCASE 2018에서 제시된 baseline 인식기의 학습절차에 따라 소리 이벤트를 검출한 결과를 비교한 것이다.

표 2

	Test set	
	DCASE 2018 CRNN baseline	분절모델 (Segmental Model)
F-Score	2.74	14.87

- [0142] 표 2와 같이, 본 발명의 검출장치(100)는 소리 이벤트 검출에 있어서, 기존의 CRNN 기반의 검출장치에 비해 향상된 성능을 나타낼 수 있다.
- [0143] 도 5 는 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 불완전 레이블 오류 검출을 설명하는데 참조되는 도이다.
- [0144] 도 5의 (a)에 도시된 바와 같이, 오디오데이터(121)에는 복수의 소리가 포함될 수 있다. 제 1 구간(P1)은 흐르는 물소리이고, 제 2 구간(P2)은 사람의 음성이 포함될 수 있다. 제 1 구간(P1)과 제2구간(P2) 사이에는 소정 길이의 묵음 구간이 포함될 수 있다.
- [0145] 도 5의 (b)에 도시된 바와 같이, 오디오데이터(121)에는 묵음 구간(P3)이 포함될 수 있다. 또한, 오디오데이터(121)에는 의미 없는 소리(Garbage)가 포함될 수 있다.
- [0146] 본 발명의 검출장치(100)는 하나의 오디오데이터(121)에 복수의 소리가 포함되거나, 묵음 구간(P3)이 포함되더라도, 세그먼트모델(분절모델)을 기반으로 오디오데이터(121)로부터 각각의 소리를 구분하여 소리 이벤트로 검출할 수 있다.
- [0147] 검출장치(100)는 오디오데이터(121)의 시작과 끝부분 그리고 소리의 연결부분에 "SIL"이라고 하는 임의의 소리 레이블(Label)을 사용할 수 있다.
- [0148] 검출장치(100)는 "SIL" 레이블을 적용하여, 하나의 오디오데이터(121)에 복수의 소리가 포함되는 경우, 오디오데이터(121)에 분류하고자 하는 타입의 소리가 아닌 경우 및 묵음 구간이 포함될 때, 세그먼트 모델링을 수행하고, 이를 바탕으로 소리 이벤트를 검출할 수 있다.
- [0149] 프로세서(110)는 임의의 소리 레이블인 SIL 레이블을 참레이블 중 하나로 설정하고, 참레이블 비교방식을 이용하여 오디오데이터(121)에 포함되는 묵음구간을 검출할 수 있다. 또한 프로세서(110)는 SIL 레이블을 이용하여 오디오데이터(121)에 포함되는 복수의 소리에 대하여 소리와 소리 사이의 묵음구간을 검출하여 상이한 소리를 구분할 수 있다.
- [0150] 다음 표 3은 SIL 레이블을 적용하는 경우와 적용하지 않은 경우에 대한 테스트 결과이다.

표 3

RNNs의 층수		Test set	Validation set	Train set
with 'SIL'	1	14.87	9.23	57.07
	2	10.49	9.26	42.44
without 'SIL'	1	6.41	4.28	7.48
	2	8.03	5.56	9.53

[0151]

- [0152] 표 3을 참조하면, 검출장치(100)는 분절모델을 기반으로 SIL 레이블을 적용하는 경우가 SIL 레이블을 적용하지 않은 경우보다 F-score가 높으므로, 성능이 향상됨을 알 수 있다.
- [0153] 학습 및 인식 시에 사용되는 오디오데이터(121)에는 우리가 분류하고자 하는 소리 외에도 많은 잡음과 묵음 구간이 존재하기 때문이다.
- [0154] 그에 따라 본 발명의 검출장치(100)는 분절모델을 기반으로, 연산량을 감소시켜 효율적으로 연산처리를 수행할 수 있고, SIL 레이블을 적용하여 소리 이벤트를 정확하게 검출할 수 있다.
- [0155] 도 6 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 분절모델 생성방법이 도시된 순서도이다.
- [0156] 도 6에 도시된 바와 같이, 검출장치(100)의 프로세서(110)는 오디오데이터(121)가 입력되면(S310), 오디오데이터(121)로부터 특징을 추출한다(S320). 이때 오디오데이터(121)는 세그먼트모델 생성을 위한 학습데이터이다.
- [0157] 프로세서(110)는 프레임 단위로 오디오데이터(121)의 특징을 추출할 수 있다. 프로세서(110)는 앞서 설명한 도 2와 같이 멜-필터링 및 로그 변환을 통해 로그-멜 필터뱅크를 산출할 수 있다.
- [0158] 프로세서(110)는 특징으로 추출된 로그-멜 필터뱅크를 인코딩한다(S330).
- [0159] 프로세서(110)는 은닉 뉴런을 가진 bi-directional LSTM(Long Short-Term Memory)과 CNN을 기반으로 인코딩할 수 있다. 프로세서(110)는 LSTM을 이용하여 추출된 특징에 대한 시간의 상관관계를 모델링하고, CNN의 풀링(pooling)을 이용하여 시간 구간의 길이를 축소함으로써 연산량을 감소시킬 수 있다.
- [0160] 프로세서(110)는 인코딩을 통해 새로운 특징 h를 생성한다.
- [0161] 프로세서(110)는 새로운 특징 h에 대하여 세그먼트 임베딩을 수행한다(S340). 프로세서(110)는 프레임 단위의 특징으로부터 세그먼트 단위의 세그먼트데이터(123)를 생성할 수 있다.
- [0162] 프로세서(110)는 세그먼트 임베딩에서, 인코딩 결과인 h로부터 각 세그먼트에 대한 임베딩 벡터를 추출한다. 프로세서(110)는 일정 시간 구간에 해당하는 임의의 세그먼트에 대하여, FNN을 이용하여 임베딩 벡터를 획득할 수 있다.
- [0163] 프로세서(110)는 세그먼트에 대한 임베딩 벡터를 획득한 후, 이를 이용하여 세그먼트에 대한 유사도 스코어를 산출한다(S350).
- [0164] 예를 들어, 프로세서(110)는 시작 시점이 t이고 세그먼트의 길이가 s이며, 소리 종류가 y인 세그먼트에 대하여, 유사한 정도에 대하여 스코어를 산출할 수 있다.
- [0165] 프로세서(110)는 세그먼트모델을 생성하는 경우, 학습을 위해 비용함수를 산출하여 적용한다(S360). 프로세서(110)는 학습을 위한 비용함수로써, Marginal Log Loss를 사용할 수 있다.
- [0166] 프로세서(110)는 입력된 오디오데이터(121)에 해당 경로를 기반으로, 각 경로를 구성하는 세그먼트에 대한 스코어를 연산하고, 비용함수를 적용한 연산결과에 대하여 반복적으로 학습을 수행한다(S370).
- [0167] 프로세서(110)는 학습 결과를 기반으로 세그먼트모델(분절모델)에 대한 파라미터를 획득할 수 있다. 프로세서(110)는 학습을 통해 획득되는 파라미터를 기반으로 세그먼트모델(분절모델)(124)을 생성할 수 있다(S380).
- [0168] 프로세서(110)는 생성된 세그먼트모델(분절모델)(124)을 메모리(120)에 저장한다.
- [0169] 도 7 은 본 발명의 일 실시예에 따른 소리 이벤트 검출장치의 분절모델 기반 소리 이벤트 검출방법이 도시된 순서도이다.
- [0170] 도 7에 도시된 바와 같이, 검출장치(100)의 프로세서(110)는 소리 이벤트를 검출할 오디오데이터(121)를 입력받는다(S410). 이때 오디오데이터(121)는 세그먼트모델을 기반으로 소리 이벤트를 검출할 데이터이다.
- [0171] 프로세서(110)는 오디오데이터(121)로부터 특징을 추출하고(S420), 인코딩을 수행한다(S430).
- [0172] 프로세서(110)는 인코딩을 통해 생성된, 새로운 특징 h에 대하여 세그먼트 임베딩을 수행한다(S440).
- [0173] 프로세서(110)는 세그먼트 임베딩에서, 인코딩 결과인 h로부터 각 세그먼트에 대한 임베딩 벡터를 추출한다. 프

로세서(110)는 일정 시간 구간에 해당하는 임의의 세그먼트에 대하여, FNN을 이용하여 임베딩 벡터를 획득할 수 있다.

[0174] 프로세서(110)는 세그먼트에 대한 임베딩 벡터를 획득한 후, 이를 이용하여 세그먼트에 대한 유사도 스코어를 산출한다(S450).

[0175] 프로세서(110)는 동적프로그래밍을 이용하여 연산되는 오디오데이터(121)의 최적 경로를 산출하여 최적의 세그먼트를 결정한다(S460).

[0176] 프로세서(110)는 앞서 설명한 도 7과 같이 유사도 스코어가 가장 큰(최대값) 최적의 경로 p^* 를 획득함으로써, 최적의 분할 정보와 그에 대응하는 소리의 종류를 확인할 수 있다.

[0177] 프로세서(110)는 스코어가 가장 큰 세그먼트를 산출하는 과정을 반복하여 오디오데이터(121)에 포함되는 소리 이벤트를 검출할 수 있다(S480).

[0178] 또한, 프로세서(110)는 묵음구간에 대한 SIL 레이블을 적용하여 오디오데이터(121)에 포함된 묵음구간을 검출하여, 소리 이벤트를 검출할 수 있다.

[0179] 프로세서(110)는 오디오데이터(121)에 포함된 묵음구간을 기준으로 전, 후 구간의 소리를 상이한 소리로 판단할 수 있다. 또한 프로세서(110)는 오디오데이터(121)의 처음과 끝에 묵음구간이 포함된 경우 이를 배제하고 나머지 구간에서 소리 이벤트를 검출할 수 있다.

[0180] 따라서 본 발명의 검출장치(100)는 분할 정보 없이도 분절모형을 기반으로 오디오데이터(121)를 세그먼트 단위로 인식하여 소리 이벤트를 검출할 수 있다.

[0181] 본 발명의 세그먼트모형(분절모형)은 오디오 특징벡터에 대한 인코딩 부분과 각 세그먼트를 대표하는 벡터를 추출하기 위한 임베딩 부분 그리고 각 세그먼트에 대한 유사도 값을 추출하기 위한 스코어링 부분들이 연결되어 있다. 세그먼트모형(분절모형)은 적절히 사용함으로써 전체적으로 엔드-투-엔드(end-to-end) 기반의 분절 기반으로 소리를 분류할 수 있고, 인식의 부자유스러움을 해소할 수 있어서 소리 이벤트 검출에 따른 성능을 향상시킬 수 있다.

[0182] 따라서 본 발명의 일 측면에 따른 분절모형 기반 소리 이벤트 검출장치 및 그 방법은, 분절모형 기반으로 오디오데이터(121)로부터 소리 단위로 신호를 추출하여, 소리 이벤트를 검출할 수 있다. 또한, 본 발명은 오디오데이터(121)에 불완전 레이블 오류가 포함된 경우에도 효과적으로 소리 이벤트를 검출할 수 있고, 검출의 정확도를 향상시킬 수 있다.

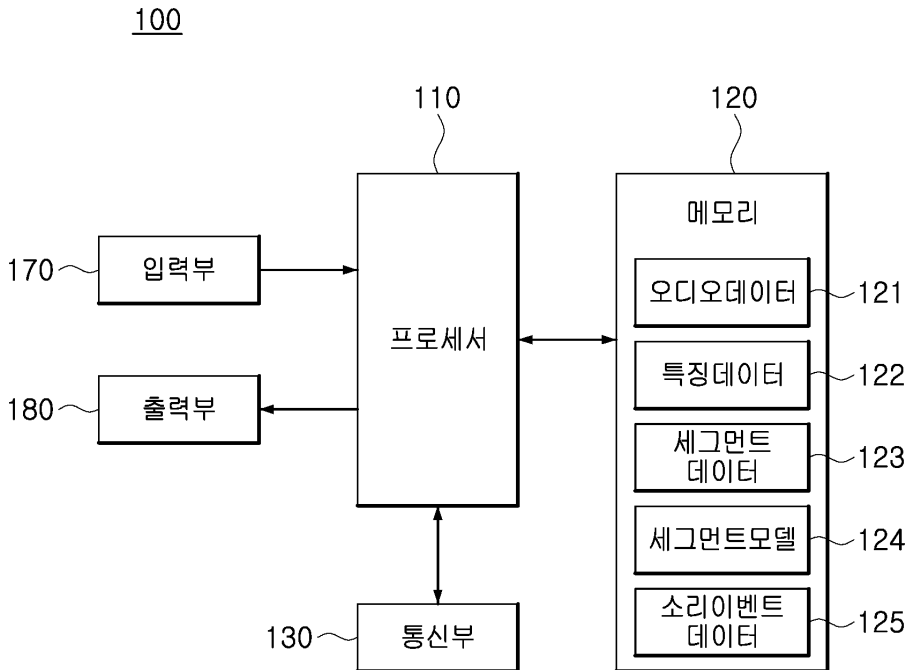
[0183] 본 발명은 도면에 도시된 실시예를 참고로 하여 설명되었으나, 이는 예시적인 것에 불과하며 당해 기술이 속하는 기술분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 아래의 특허청구범위에 의하여 정해져야할 것이다.

부호의 설명

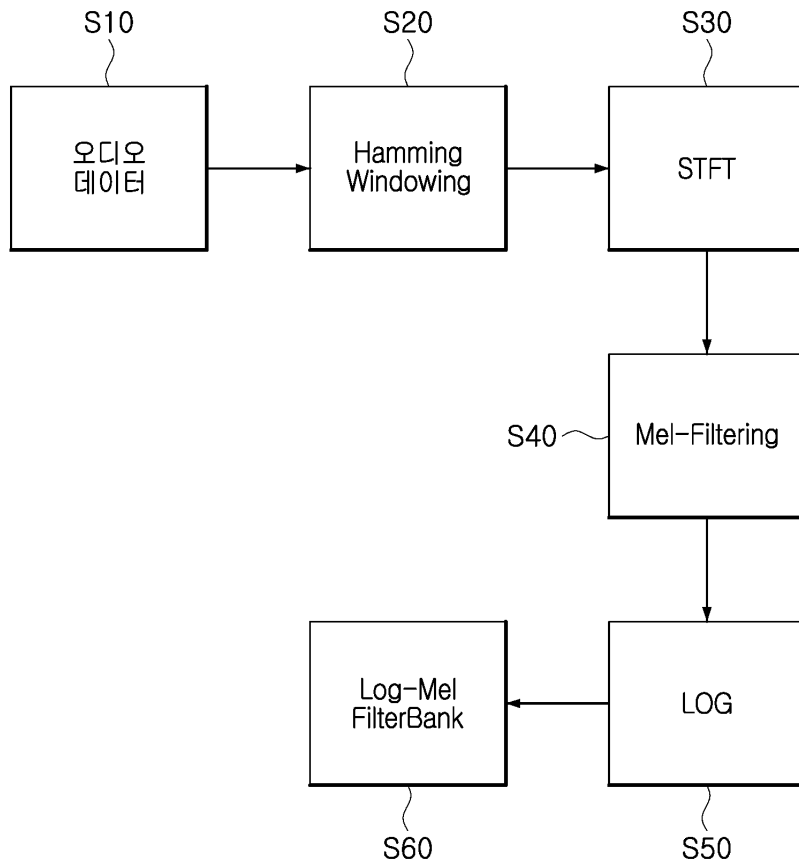
[0184]	100: 검출장치	
	110: 프로세서	120: 메모리
	121: 오디오데이터	122: 특징데이터
	123: 세그먼트데이터	124: 세그먼트모형(분절모형)
	125: 소리이벤트데이터	
	130: 통신부	
	170: 입력부	180: 출력부

도면

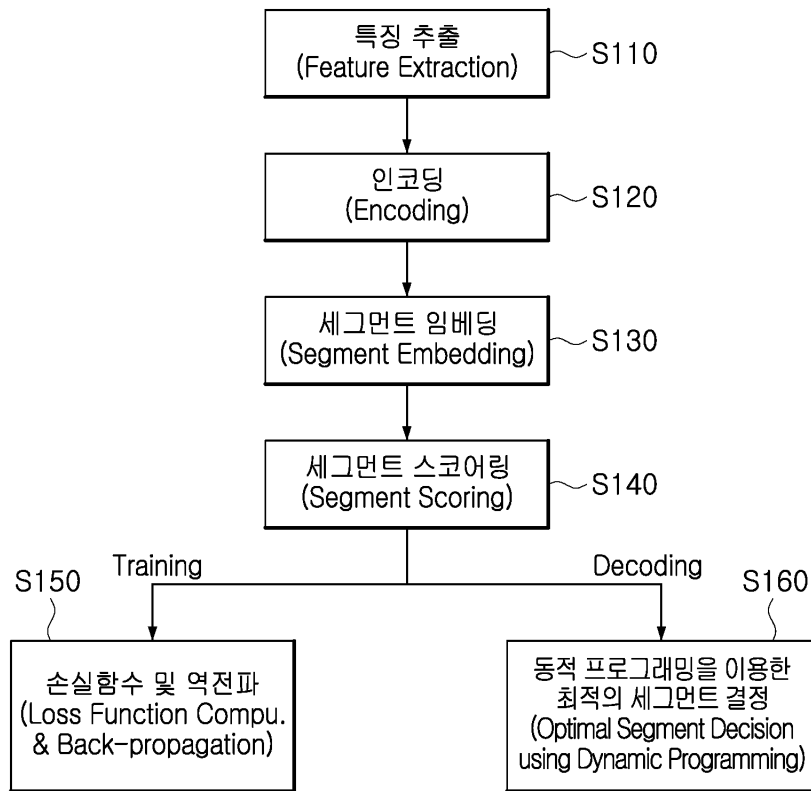
도면1



도면2



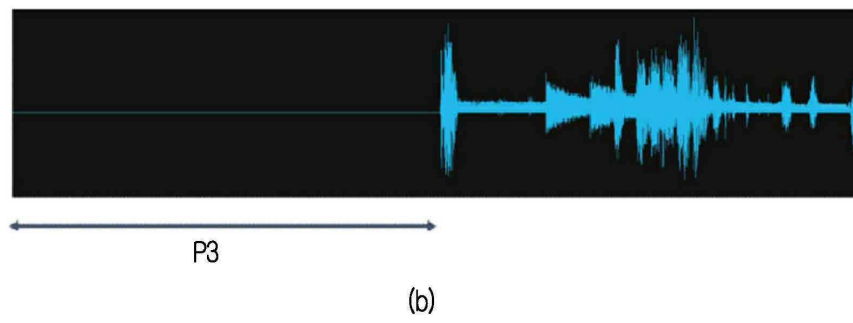
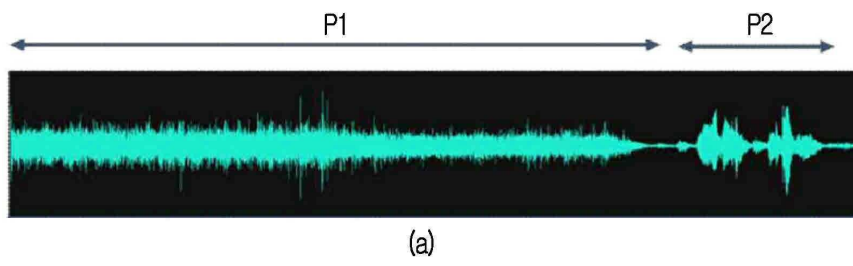
도면3



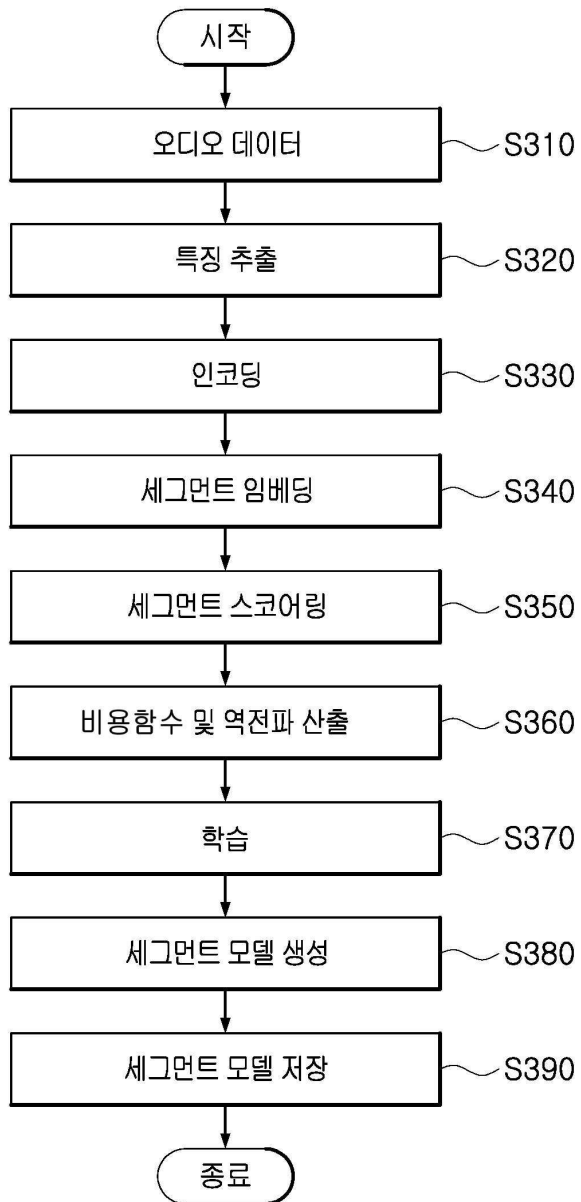
도면4

	Train	Test	Validation
No. of clips	2548	692	1168
Clip lenght	10 sec.		
Classes(10)	Speech, Dog, Cat, Alarm bell ring Dishes, Frying, Blender, Running water, Vacuum cleaner, Electric shaver toothbrush		

도면5



도면6



도면7

