



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0074548
(43) 공개일자 2025년05월27일

- (51) 국제특허분류(Int. Cl.)
G10L 13/08 (2006.01) G06F 16/63 (2019.01)
G06F 40/279 (2020.01) G06N 3/0455 (2023.01)
G06N 3/0475 (2023.01) G06N 3/096 (2023.01)
G10L 15/06 (2006.01) G10L 15/26 (2006.01)
G10L 21/0208 (2013.01) G10L 21/055 (2013.01)
- (52) CPC특허분류
G10L 13/08 (2013.01)
G06F 16/63 (2019.01)
- (21) 출원번호 10-2024-0026317
(22) 출원일자 2024년02월23일
심사청구일자 2024년02월23일
- (30) 우선권주장
1020230160225 2023년11월20일 대한민국(KR)
- (71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
- (72) 발명자
김태환
울산광역시 울주군 언양읍 유니스트길 50
정석훈
울산광역시 울주군 언양읍 유니스트길 50
- (74) 대리인
특허법인 무한

전체 청구항 수 : 총 22 항

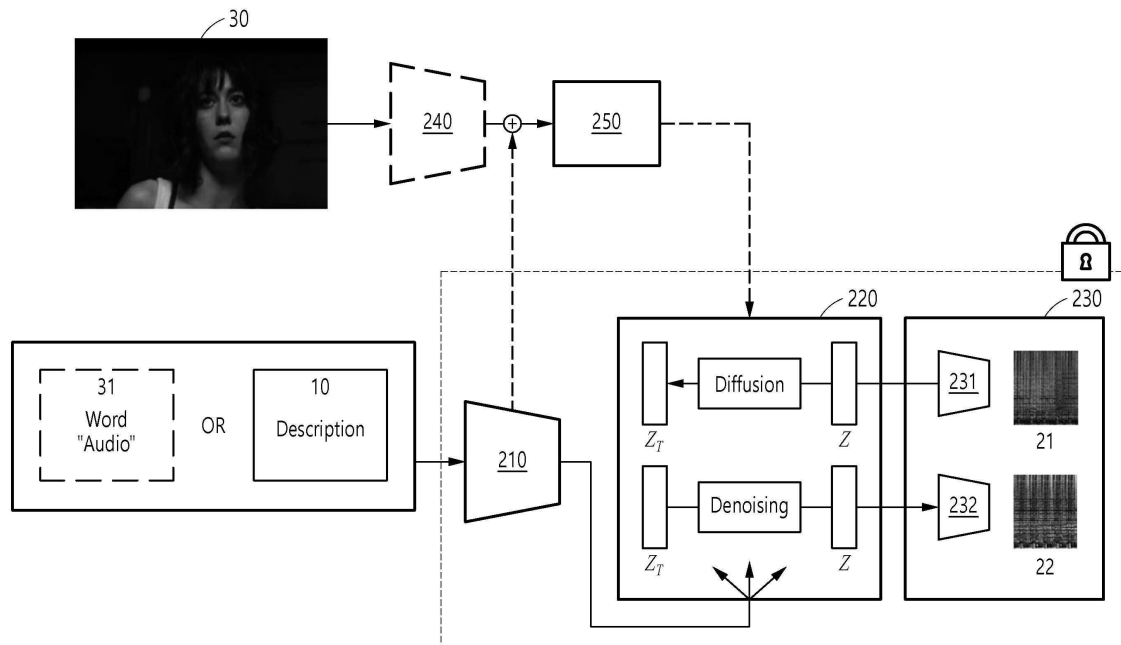
(54) 발명의 명칭 이미지 및 텍스트 기반으로 오디오를 생성하는 오디오 생성 모델을 트레이닝시키는 방법 및 장치

(57) 요약

일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 사전-트레이닝된(pre-trained) 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 상기 오디오 생성 모델의 텍스트 조건 형태로 변환하는 단계, 상기 변환된 텍스트 조건 형태 및 상기 트레

(뒷면에 계속)

대표도



이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 기반 손실(loss)을 산출하는 단계, 상기 산출된 텍스트 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터(parameter)를 결정하는 단계, 상기 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 상기 오디오 생성 모델의 이미지 조건 형태로 변환하는 단계, 상기 텍스트 기반 오디오 생성 파라미터, 상기 변환된 이미지 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출하는 단계 및 상기 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하는 단계를 포함할 수 있다.

(52) CPC특허분류

G06F 40/279 (2020.01)

G06N 3/0455 (2023.01)

G06N 3/0475 (2023.01)

G06N 3/096 (2023.01)

G10L 15/063 (2013.01)

G10L 15/26 (2013.01)

G10L 21/0208 (2013.01)

G10L 21/055 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193081
과제번호	2022-0-00608-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	(1세부) 인간과 교감하는 멀티모달 인터랙션 인공지능 기술
과제수행기관명	한국전자기술연구원
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193622
과제번호	2021-0-02068-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능 혁신 허브 연구 개발
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

프로세서에 의해 수행되는 오디오 생성 모델을 트레이닝시키는 방법에 있어서,

사전-트레이닝(pre-trained) 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 상기 오디오 생성 모델의 텍스트 조건 형태로 변환하는 단계;

상기 변환된 텍스트 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 기반 손실(loss)을 산출하는 단계;

상기 산출된 텍스트 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터(parameter)를 결정하는 단계;

상기 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 상기 오디오 생성 모델의 이미지 조건 형태로 변환하는 단계;

상기 텍스트 기반 오디오 생성 파라미터, 상기 변환된 이미지 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출하는 단계; 및

상기 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하는 단계

를 포함하는 방법.

청구항 2

제1항에 있어서,

상기 텍스트 조건 형태로 변환하는 단계는,

상기 텍스트 인코더에 상기 텍스트 데이터를 입력하여 텍스트 임베딩(embedding)을 추출하는 단계

를 포함하는 방법.

청구항 3

제2항에 있어서,

상기 텍스트 기반 손실을 산출하는 단계는,

상기 오디오 데이터를 참값(Ground Truth)으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간(pixel space)의 고 차원 참값을 잠재 공간(latent space)의 저 차원 참값으로 변환하는 단계;

상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전(diffusion)하여 노이즈가 삽입된 저 차원 참값을 생성하는 단계;

상기 텍스트 임베딩을 조건으로 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이징(denoising)하여 노이즈가 제거된 저 차원 예측값을 생성하는 단계;

상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하는 단계; 및

상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 기반 손실을 산출하는 단계

를 포함하는 방법.

청구항 4

제1항에 있어서,

상기 이미지 조건 형태로 변환하는 단계는,

상기 이미지 인코더에 상기 이미지 데이터를 입력하여 이미지 임베딩을 추출하는 단계
를 포함하는 방법.

청구항 5

제4항에 있어서,

상기 텍스트 및 이미지 기반 손실을 산출하는 단계는,

상기 오디오 데이터를 참값으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환하는 단계;

상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성하는 단계;

상기 이미지 임베딩 및 텍스트 임베딩을 조건으로 상기 텍스트 기반 오디오 생성 파라미터가 적용된 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이즈하여 노이즈가 제거된 저 차원 예측값을 생성하는 단계;

상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하는 단계; 및

상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 및 이미지 기반 손실을 산출하는 단계
를 포함하는 방법.

청구항 6

제1항에 있어서,

상기 텍스트 기반 오디오 생성 파라미터를 결정한 이후, 상기 트레이닝 데이터 세트 중 이미지 데이터로부터 하나 이상의 단어들을 추출하는 단계;

사전-트레이닝된 텍스트 인코더를 통해 상기 추출된 하나 이상의 단어들에서 이미지 기반 텍스트 임베딩을 추출하는 단계;

상기 추출된 이미지 기반 텍스트 임베딩, 상기 텍스트 기반 오디오 생성 파라미터 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 이미지 기반 손실을 산출하는 단계; 및

상기 산출된 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 이미지 기반 오디오 생성 파라미터를 결정하는 단계;

를 더 포함하는 방법.

청구항 7

제1항에 있어서,

상기 오디오 생성 모델을 트레이닝시키기 이전에, 상기 트레이닝 데이터 세트를 생성하는 단계

를 더 포함하는 방법.

청구항 8

제7항에 있어서,

상기 트레이닝 데이터 세트를 생성하는 단계는,

영상 데이터로부터 제1 지속 시간의 복수의 제1 클립 데이터들을 생성하는 단계;

상기 영상 데이터로부터 상기 제1 지속 시간보다 짧은 시간 길이인 제2 지속 시간의 복수의 제2 클립 데이터들을 생성하는 단계;

상기 제1 지속 시간에 대응되는 시간 길이가 되도록 상기 복수의 제2 클립 데이터들을 연결시키는 단계;

상기 복수의 제1 클립 데이터들과 상기 연결된 복수의 제2 클립 데이터들을 병합시켜 복수의 원시 클립 데이터를 생성하는 단계; 및

상기 복수의 원시 클립 데이터로부터 추출된 이미지 데이터 및 음성 분리 오디오 데이터와 이에 기반한 텍스트 데이터에 기초하여 복수의 상기 트레이닝 데이터 세트를 생성하는 단계

를 포함하는 방법.

청구항 9

제8항에 있어서,

상기 복수의 트레이닝 데이터 세트를 생성하는 단계는,

상기 원시 클립 데이터로부터 비디오 프레임들을 추출하고, 상기 추출된 비디오 프레임들을 클러스터링하여 하나 이상의 상기 이미지 데이터를 추출하는 단계;

상기 원시 클립 데이터로부터 원시 오디오를 추출하고, 상기 추출된 원시 오디오에서 음성(speech)을 분리(decoupling)하여 상기 음성 분리 오디오 데이터를 생성하는 단계; 및

상기 이미지 데이터와 상기 음성 분리 오디오 데이터에 상기 복수의 제1 클립 데이터들 및 상기 복수의 제2 클립 데이터들 생성 시 제공되는 텍스트 데이터를 대응시켜 오디오 데이터, 이미지 데이터 및 텍스트 데이터의 쌍의 상기 트레이닝 데이터 세트를 생성하는 단계

를 포함하는 방법.

청구항 10

제9항에 있어서,

상기 이미지 데이터를 추출하는 단계는,

상기 비디오 프레임들의 이미지 핑거프린트(fingerprint)의 유사도에 기초하여 이미지를 제거하고 남은 이미지를 클러스터링하여 상기 이미지 데이터로 구성하는 단계

를 포함하는 방법.

청구항 11

제9항에 있어서,

상기 트레이닝 데이터를 생성하는 단계는,

상기 음성 분리 오디오 데이터를 생성한 이후, 상기 음성 분리 오디오 데이터를 스펙트로그램(Spectrogram) 이미지로 변환하는 단계

를 더 포함하는 방법.

청구항 12

사전-트레이닝된 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 오디오 생성 모델의 텍스트 조건 형태로 변환하고, 상기 변환된 텍스트 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 기반 손실을 산출하고, 상기 산출된 텍스트 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터를 결정하며,

상기 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 상기 오디오 생성 모델의 이미지 조건 형태로 변환하고, 상기 텍스트 기반 오디오 생성 파라미터, 상기 변환된 이미지 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출하고, 상기 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하는 프로세서

를 포함하는 장치.

청구항 13

제12항에 있어서,

상기 프로세서는,

상기 텍스트 인코더에 상기 텍스트 데이터를 입력하여 텍스트 임베딩을 추출하는,

장치.

청구항 14

제13항에 있어서,

상기 프로세서는,

상기 오디오 데이터를 참값으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환하고, 상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성하고, 상기 텍스트 임베딩을 조건으로 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이즈하여 노이즈가 제거된 저 차원 예측값을 생성하고, 상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하며, 상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 기반 손실을 산출하는,

장치.

청구항 15

제12항에 있어서,

상기 프로세서는,

상기 이미지 인코더에 상기 이미지 데이터를 입력하여 이미지 임베딩을 추출하는,

장치.

청구항 16

제15항에 있어서,

상기 프로세서는,

상기 오디오 데이터를 참값으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환하고, 상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성하고, 상기 이미지 임베딩 및 텍스트 임베딩을 조건으로 상기 텍스트 기반 오디오 생성 파라미터가 적용된 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이징하여 노이즈가 제거된 저 차원 예측값을 생성하고, 상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하며, 상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 및 이미지 기반 손실을 산출하는,

장치.

청구항 17

제12항에 있어서,

상기 프로세서는,

상기 텍스트 기반 오디오 생성 파라미터를 결정한 이후, 상기 트레이닝 데이터 세트 중 이미지 데이터로부터 하나 이상의 단어들을 추출하고, 사전-트레이닝된 텍스트 인코더를 통해 상기 추출된 하나 이상의 단어들에서 이미지 기반 텍스트 임베딩을 추출하고, 상기 추출된 이미지 기반 텍스트 임베딩, 상기 텍스트 기반 오디오 생성 파라미터 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 이미지 기반 손실을 산출하며, 상기 산출된 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 이미지 기반 오디오 생성 파라미터를 결정하는,

장치.

청구항 18

제12항에 있어서,

상기 프로세서는,

상기 오디오 생성 모델을 트레이닝시키기 이전에, 상기 트레이닝 데이터 세트를 생성하는,

장치.

청구항 19

제18항에 있어서,

상기 프로세서는,

영상 데이터로부터 제1 지속 시간의 복수의 제1 클립 데이터들을 생성하고, 상기 영상 데이터로부터 상기 제1 지속 시간보다 짧은 시간 길이인 제2 지속 시간의 복수의 제2 클립 데이터들을 생성하고, 상기 제1 지속 시간에 대응되는 시간 길이가 되도록 상기 복수의 제2 클립 데이터들을 연결시키고, 상기 복수의 제1 클립 데이터들과 상기 연결된 복수의 제2 클립 데이터들을 병합시켜 복수의 원시 클립 데이터를 생성하며, 상기 복수의 원시 클립 데이터로부터 추출된 이미지 데이터 및 음성 분리 오디오 데이터와 이에 기반한 텍스트 데이터에 기초하여 복수의 상기 트레이닝 데이터 세트를 생성하는,

장치.

청구항 20

제19항에 있어서,

상기 프로세서는,

상기 원시 클립 데이터로부터 비디오 프레임들을 추출하고, 상기 추출된 비디오 프레임들을 클러스터링하여 하나 이상의 상기 이미지 데이터를 추출하고, 상기 원시 클립 데이터로부터 원시 오디오를 추출하고, 상기 추출된 원시 오디오에서 음성을 분리하여 상기 음성 분리 오디오 데이터를 생성하고, 상기 이미지 데이터와 상기 음성 분리 오디오 데이터에 상기 복수의 제1 클립 데이터들 및 상기 복수의 제2 클립 데이터들 생성 시 제공되는 텍스트 데이터를 대응시켜 오디오 데이터, 이미지 데이터 및 텍스트 데이터의 쌍의 상기 트레이닝 데이터 세트를 생성하는,

장치.

청구항 21

제20항에 있어서,

상기 프로세서는,

상기 비디오 프레임들의 이미지 핑거프린트의 유사도에 기초하여 이미지를 제거하고 남은 이미지를 클러스터링하여 상기 이미지 데이터로 구성하는,

장치.

청구항 22

제20항에 있어서,

상기 프로세서는,

상기 음성 분리 오디오 데이터를 생성한 이후, 상기 음성 분리 오디오 데이터를 스펙트로그램 이미지로 변환하는,

장치.

발명의 설명

기술 분야

[0001]

이하, 이미지 시퀀스, 텍스트 또는 둘 다에 기초하여 배경 음악 및 효과 음을 생성하는 오디오 생성 모델을 트레이닝시키는 것에 관한 기술이 제공된다.

배경 기술

[0003]

생성형(Generative) AI는 대규모 데이터 세트를 기반으로 트레이닝된 딥러닝 모델을 사용하여 새로운 콘텐츠를 생성하는 일종의 인공지능 기술을 의미한다. 생성형 AI를 통해 생성되는 새로운 콘텐츠는 대화(chat), 이야기(story), 이미지(image), 비디오(video), 음악(music) 등을 포함할 수 있다. 생성형 AI는 방대한 경험 데이터 베이스에 인공지능을 이용하여 새롭고 독창적인 정보를 생성할 수 있어, 금융 서비스, 의료와 생명 과학, 자동차와 제조, 미디어와 엔터테인먼트, 전기통신 및 에너지 등 다양한 분야에 적용될 수 있다. 예를 들어, 생성형

AI는 약물의 효과를 최적화한 화합물을 생성하여 신약을 개발할 수 있고, 이미지 또는 텍스트 입력을 기반으로 이미지를 생성할 수 있다.

- [0004] 위에서 설명한 배경기술은 발명자가 본원의 개시 내용을 도출하는 과정에서 보유하거나 습득한 것으로서, 반드시 본 출원 전에 일반 공중에 공개된 공지기술이라고 할 수는 없다.

선행기술문헌

특허문헌

- [0006] (특허문헌 0001) 한국 등록 특허공보 10-2138248(2020년 07월 21일 공고)에는 온라인 채팅의 배경음악생성방법 및 그 장치가 제시된다.
- (특허문헌 0002) 한국 등록 특허공보 10-1896193(2018년 08월 31일 공고)에는 이미지를 음악으로 변환하는 방법이 제시된다.

발명의 내용

해결하려는 과제

- [0007] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 텍스트 및/또는 이미지를 조건으로 트레이닝된 오디오 생성 모델을 이용하여 자연어를 처리함에 따라 텍스트, 이미지 및 영상의 스토리를 분석하고, 분석된 스토리를 이용하여 오디오를 생성할 수 있다.
- [0008] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 오디오 생성 모델을 트레이닝하여 텍스트 기반 오디오 생성 파라미터를 우선적으로 결정하고, 결정된 텍스트 기반 오디오 생성 파라미터에 기초하여 오디오 생성 모델을 추가적으로 트레이닝함에 따라 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정할 수 있다.
- [0009] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 동일한 영상 데이터에 각기 다른 방식을 이용하여 복수의 제1 클립 데이터들과 복수의 제2 클립 데이터들을 추출하고 이를 병합하여 트레이닝 데이터 세트를 생성할 수 있다.
- [0010] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 트레이닝 수행 전에 음성이 분리된 오디오와 이미지만을 설명하는 텍스트(예: 캡션)가 아닌 스토리를 포함하는 텍스트(예: 설명)를 포함하는 트레이닝 데이터 세트를 생성할 수 있다.
- [0011] 다만, 기술적 과제는 상술한 기술적 과제들로 한정되는 것은 아니며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

- [0013] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법은 사전-트레이닝된(pre-trained) 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 상기 오디오 생성 모델의 텍스트 조건 형태로 변환하는 단계, 상기 변환된 텍스트 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 기반 손실(loss)을 산출하는 단계, 상기 산출된 텍스트 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터(parameter)를 결정하는 단계, 상기 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 상기 오디오 생성 모델의 이미지 조건 형태로 변환하는 단계, 상기 텍스트 기반 오디오 생성 파라미터, 상기 변환된 이미지 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출하는 단계 및 상기 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하는 단계를 포함할 수 있다.
- [0014] 상기 텍스트 조건 형태로 변환하는 단계는, 상기 텍스트 인코더에 상기 텍스트 데이터를 입력하여 텍스트 임베

딩(embedding)을 추출하는 단계를 포함할 수 있다.

- [0015] 상기 텍스트 기반 손실을 산출하는 단계는, 상기 오디오 데이터를 참값(Ground Truth)으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간(pixel space)의 고 차원 참값을 잠재 공간(latent space)의 저 차원 참값으로 변환하는 단계, 상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전(diffusion)하여 노이즈가 삽입된 저 차원 참값을 생성하는 단계, 상기 텍스트 임베딩을 조건으로 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이징(denoising)하여 노이즈가 제거된 저 차원 예측값을 생성하는 단계, 상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하는 단계 및 상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 기반 손실을 산출하는 단계를 포함할 수 있다.
- [0016] 상기 이미지 조건 형태로 변환하는 단계는, 상기 이미지 인코더에 상기 이미지 데이터를 입력하여 이미지 임베딩을 추출하는 단계를 포함할 수 있다.
- [0017] 상기 텍스트 및 이미지 기반 손실을 산출하는 단계는, 상기 오디오 데이터를 참값으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환하는 단계, 상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성하는 단계, 상기 이미지 임베딩 및 텍스트 임베딩을 조건으로 상기 텍스트 기반 오디오 생성 파라미터가 적용된 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이징하여 노이즈가 제거된 저 차원 예측값을 생성하는 단계, 상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하는 단계 및 상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 및 이미지 기반 손실을 산출하는 단계를 포함할 수 있다.
- [0018] 상기 방법은 상기 텍스트 기반 오디오 생성 파라미터를 결정한 이후, 상기 트레이닝 데이터 세트 중 이미지 데이터로부터 하나 이상의 단어들을 추출하는 단계, 사전-트레이닝된 텍스트 인코더를 통해 상기 추출된 하나 이상의 단어들에서 이미지 기반 텍스트 임베딩을 추출하는 단계, 상기 추출된 이미지 기반 텍스트 임베딩, 상기 텍스트 기반 오디오 생성 파라미터 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 이미지 기반 손실을 산출하는 단계 및 상기 산출된 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 이미지 기반 오디오 생성 파라미터를 결정하는 단계를 더 포함할 수 있다.
- [0019] 상기 방법은 상기 오디오 생성 모델을 트레이닝시키기 이전에, 상기 트레이닝 데이터 세트를 생성하는 단계를 더 포함할 수 있다.
- [0020] 상기 트레이닝 데이터 세트를 생성하는 단계는, 영상 데이터로부터 제1 지속 시간의 복수의 제1 클립 데이터들을 생성하는 단계, 상기 영상 데이터로부터 상기 제1 지속 시간보다 짧은 시간 길이인 제2 지속 시간의 복수의 제2 클립 데이터들을 생성하는 단계, 상기 제1 지속 시간에 대응되는 시간 길이가 되도록 상기 복수의 제2 클립 데이터들을 연결시키는 단계, 상기 복수의 제1 클립 데이터와 상기 연결된 복수의 제2 클립 데이터들을 병합시켜 복수의 원시 클립 데이터를 생성하는 단계 및 상기 복수의 원시 클립 데이터로부터 추출된 이미지 데이터 및 음성 분리 오디오 데이터와 이에 기반한 텍스트 데이터에 기초하여 복수의 상기 트레이닝 데이터 세트를 생성하는 단계를 포함할 수 있다.
- [0021] 상기 복수의 트레이닝 데이터 세트를 생성하는 단계는, 상기 원시 클립 데이터로부터 비디오 프레임들을 추출하고, 상기 추출된 비디오 프레임들을 클러스터링하여 하나 이상의 상기 이미지 데이터를 추출하는 단계, 상기 원시 클립 데이터로부터 원시 오디오를 추출하고, 상기 추출된 원시 오디오에서 음성(speech)을 분리(decoupling)하여 상기 음성 분리 오디오 데이터를 생성하는 단계 및 상기 이미지 데이터와 상기 음성 분리 오디오 데이터에 상기 복수의 제1 클립 데이터 및 상기 복수의 제2 클립 데이터를 생성 시 제공되는 텍스트 데이터를 대응시켜 오디오 데이터, 이미지 데이터 및 텍스트 데이터의 쌍의 상기 트레이닝 데이터 세트를 생성하는 단계를 포함할 수 있다.
- [0022] 상기 이미지 데이터를 추출하는 단계는, 상기 비디오 프레임들의 이미지 핑거프린트(fingerprint)의 유사도에 기초하여 이미지를 제거하고 남은 이미지를 클러스터링하여 상기 이미지 데이터로 구성하는 단계를 포함할 수 있다.
- [0023] 상기 트레이닝 데이터를 생성하는 단계는, 상기 음성 분리 오디오 데이터를 생성한 이후, 상기 음성 분리 오디오 데이터를 스펙트로그램(Spectrogram) 이미지로 변환하는 단계를 더 포함할 수 있다.
- [0024] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 장치는 사전-트레이닝된 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 오디오 생성 모델의 텍스트 조건 형태로 변환하고, 상기 변환된 텍스트 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이

터에 기초하여, 상기 오디오 생성 모델의 텍스트 기반 손실을 산출하고, 상기 산출된 텍스트 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터를 결정하며, 상기 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 상기 오디오 생성 모델의 이미지 조건 형태로 변환하고, 상기 텍스트 기반 오디오 생성 파라미터, 상기 변환된 이미지 조건 형태 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출하고, 상기 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하는 프로세서를 포함할 수 있다.

[0025] 상기 프로세서는, 상기 텍스트 인코더에 상기 텍스트 데이터를 입력하여 텍스트 임베딩을 추출할 수 있다.

[0026] 상기 프로세서는, 상기 오디오 데이터를 참값으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환하고, 상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성하고, 상기 텍스트 임베딩을 조건으로 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이징하여 노이즈가 제거된 저 차원 예측값을 생성하고, 상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하며, 상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 기반 손실을 산출할 수 있다.

[0027] 상기 프로세서는, 상기 이미지 인코더에 상기 이미지 데이터를 입력하여 이미지 임베딩을 추출할 수 있다.

[0028] 상기 프로세서는, 상기 오디오 데이터를 참값으로 하여, 상기 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환하고, 상기 변환된 저 차원 참값을 상기 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성하고, 상기 이미지 임베딩 및 텍스트 임베딩을 조건으로 상기 텍스트 기반 오디오 생성 파라미터가 적용된 상기 오디오 생성 모델을 통해 상기 노이즈가 삽입된 저 차원 참값을 디노이징하여 노이즈가 제거된 저 차원 예측값을 생성하고, 상기 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환하며, 상기 고 차원 참값과 상기 고 차원 예측값 간의 텍스트 및 이미지 기반 손실을 산출할 수 있다.

[0029] 상기 프로세서는, 상기 텍스트 기반 오디오 생성 파라미터를 결정한 이후, 상기 트레이닝 데이터 세트 중 이미지 데이터로부터 하나 이상의 단어들을 추출하고, 사전-트레이닝된 텍스트 인코더를 통해 상기 추출된 하나 이상의 단어들에서 이미지 기반 텍스트 임베딩을 추출하고, 상기 추출된 이미지 기반 텍스트 임베딩, 상기 텍스트 기반 오디오 생성 파라미터 및 상기 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 상기 오디오 생성 모델의 이미지 기반 손실을 산출하며, 상기 산출된 이미지 기반 손실이 최소화되도록 상기 오디오 생성 모델을 트레이닝시켜 이미지 기반 오디오 생성 파라미터를 결정할 수 있다.

[0030] 상기 프로세서는, 상기 오디오 생성 모델을 트레이닝시키기 이전에, 상기 트레이닝 데이터 세트를 생성할 수 있다.

[0031] 상기 프로세서는, 영상 데이터로부터 제1 지속 시간의 복수의 제1 클립 데이터들을 생성하고, 상기 영상 데이터로부터 상기 제1 지속 시간보다 짧은 시간 길이인 제2 지속 시간의 복수의 제2 클립 데이터들을 생성하고, 상기 제1 지속 시간에 대응되는 시간 길이가 되도록 상기 복수의 제2 클립 데이터들을 연결시키고, 상기 복수의 제1 클립 데이터들과 상기 연결된 복수의 제2 클립 데이터들을 병합시켜 복수의 원시 클립 데이터를 생성하며, 상기 복수의 원시 클립 데이터로부터 추출된 이미지 데이터 및 음성 분리 오디오 데이터와 이에 기반한 텍스트 데이터에 기초하여 복수의 상기 트레이닝 데이터 세트를 생성할 수 있다.

[0032] 상기 프로세서는, 상기 원시 클립 데이터로부터 비디오 프레임들을 추출하고, 상기 추출된 비디오 프레임들을 클러스터링하여 하나 이상의 상기 이미지 데이터를 추출하고, 상기 원시 클립 데이터로부터 원시 오디오를 추출하고, 상기 추출된 원시 오디오에서 음성을 분리하여 상기 음성 분리 오디오 데이터를 생성하고, 상기 이미지 데이터와 상기 음성 분리 오디오 데이터에 상기 복수의 제1 클립 데이터들 및 상기 복수의 제2 클립 데이터들 생성 시 제공되는 텍스트 데이터를 대응시켜 오디오 데이터, 이미지 데이터 및 텍스트 데이터의 쌍의 상기 트레이닝 데이터 세트를 생성할 수 있다.

[0033] 상기 프로세서는, 상기 비디오 프레임들의 이미지 핑거프린트의 유사도에 기초하여 이미지를 제거하고 남은 이미지를 클러스터링하여 상기 이미지 데이터로 구성할 수 있다.

[0034] 상기 프로세서는, 상기 음성 분리 오디오 데이터를 생성한 이후, 상기 음성 분리 오디오 데이터를 스펙트로그램 이미지로 변환할 수 있다.

발명의 효과

- [0036] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 텍스트 및/또는 이미지를 조건으로 트레이닝된 오디오 생성 모델을 이용하여 자연어를 처리함에 따라 텍스트, 이미지 및 영상의 스토리를 분석하고, 분석된 스토리를 이용하여 더욱 사실적인 오디오를 생성할 수 있고 이를 통해 오디오 생성에 들어가는 비용(예: 인건비)이 절감될 수 있다.
- [0037] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 오디오 생성 모델을 트레이닝하여 텍스트 기반 오디오 생성 파라미터를 우선적으로 결정하고, 결정된 텍스트 기반 오디오 생성 파라미터에 기초하여 오디오 생성 모델을 추가적으로 트레이닝함에 따라 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하므로 전체적인 트레이닝 연산이 간단해질 수 있고 전체적인 트레이닝 과정을 빠르게 진행할 수 있다.
- [0038] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 동일한 영상 데이터에 각기 다른 방식을 이용하여 복수의 제1 클립 데이터들과 복수의 제2 클립 데이터들을 추출하고 이를 병합하여 트레이닝 데이터 세트를 생성하므로, 오디오 생성 모델이 상호 보완적으로 트레이닝될 수 있다.
- [0039] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법 및 장치는 트레이닝 수행 전에 음성이 분리된 오디오와 이미지만을 설명하는 텍스트(예: 캡션)가 아닌 스토리를 포함하는 텍스트(예: 설명)를 포함하는 트레이닝 데이터 세트를 생성함에 따라 이를 이용하여 오디오 생성 모델이 스토리에 기반한 배경음(Background sound)을 생성할 수 있도록 트레이닝시킬 수 있다.

도면의 간단한 설명

- [0041] 도 1은 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 장치의 블록도를 도시한다.
- 도 2는 일 실시예에 따른 오디오 생성 모델의 아키텍처를 예시적으로 나타낸다.
- 도 3은 일 실시예에 따른 오디오 생성 모델의 트레이닝을 위한 트레이닝 데이터 세트를 예시적으로 나타낸다.
- 도 4는 일 실시예에 따른 오디오 생성 모델의 트레이닝을 위한 트레이닝 데이터 세트의 오디오 데이터, 이미지 데이터 및 텍스트 데이터가 생성되는 과정을 예시적으로 나타낸다.
- 도 5는 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법의 흐름도를 도시한다.
- 도 6은 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법으로 트레이닝된 오디오 생성 모델과 다른 음악 생성 모델의 FAD(Frechet Audio Distance)기반 평가 결과를 나타낸다.
- 도 7은 일 실시예에 따른 트레이닝 데이터 세트로 트레이닝된 검색 모델의 아키텍처를 예시적으로 나타낸다.

발명을 실시하기 위한 구체적인 내용

- [0042] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0043] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0044] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.
- [0045] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로써 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

- [0046] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0047] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.
- [0048] 자연어 처리(Natural Language Processing, NLP)는 인간이 일상적으로 사용하는 자연어를 기계적으로 처리하는 기술을 의미할 수 있다. 이러한 자연어 처리는 자연어 분석, 자연어 이해, 자연어 생성 등의 기술을 포함할 수 있다. 자연어 분석은 자연어의 형태소, 통사, 의미 및 화용을 분석하는 것을 의미할 수 있고, 자연어 이해는 컴퓨터가 자연어로 주어진 입력에 따라 동작하게 하는 것을 의미할 수 있다. 또한, 자연어 생성은 특정 내용을 사람이 이해할 수 있는 자연어로 변환하는 것을 의미할 수 있다. 이러한 자연어 처리는 인공지능(AI)의 주요한 개발 분야 중 하나로 머신 러닝(Machine Learning, ML) 및/또는 딥 러닝(Deep learning) 기술 등을 이용하여 수행될 수 있다. 예를 들어, 챗봇은 대화 사용자의 입력되는 대화를 이해하고 이에 적절한 답변을 제공하고 있으며, 음성 인식 플랫폼은 사용자의 음성이 입력되면 입력된 음성을 텍스트로 변환하여 사용자에게 제공할 수 있다.
- [0049] 한편, 배경 음악 및 효과음과 같이 음성이 없는 오디오는 영상, 이미지 및 텍스트 등으로 구성된 콘텐츠(예: 영화, 애니메이션, 게임 및 소셜 등)가 제공되는 경우에 콘텐츠의 몰입을 방해하지 않으며 콘텐츠의 이해를 돕는 중요한 요소에 해당될 수 있다. 콘텐츠는 일반적으로 맥락이 있는 스토리로 구성되어, 콘텐츠에 제공되는 음성이 없는 오디오 또한 해당 스토리를 따르는 것이 필요하다.
- [0050] 이에 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 장치는 텍스트 및/또는 이미지를 조건으로 트레이닝된 오디오 생성 모델을 이용하여 자연어를 처리함에 따라 텍스트, 이미지 및 영상의 스토리를 분석하고, 분석된 스토리를 이용하여 더욱 사실적인 오디오를 생성할 수 있다. 이를 통해 오디오 생성에 들어가는 인건비, 장비 대여료와 같은 부대비용 또한 절감할 수 있다.
- [0051] 도 1은 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 장치의 블록도를 도시한다.
- [0052] 장치(100)는 프로세서(110) 및 메모리(120)를 포함할 수 있다.
- [0053] 프로세서(110)는 사전-트레이닝된(pre-trained) 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 오디오 생성 모델의 텍스트 조건 형태로 변환할 수 있다. 프로세서(110)는 변환된 텍스트 조건 형태 및 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 오디오 생성 모델의 텍스트 기반 손실(loss)을 산출할 수 있다. 프로세서(110)는 산출된 텍스트 기반 손실이 최소화되도록 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터(parameter)를 결정할 수 있다. 이에 프로세서(110)는 텍스트 기반 오디오 생성 파라미터를 통해 입력된 텍스트에 기초하여 오디오를 생성할 수 있다. 또한, 오디오 생성 모델을 트레이닝시키는 트레이닝 데이터 세트는 도 3에서 상세히 후술한다.
- [0054] 또한, 프로세서(110)는 결정된 텍스트 기반 오디오 생성 파라미터에 기초하여 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 우선, 프로세서(110)는 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 오디오 생성 모델의 이미지 조건 형태로 변환할 수 있다. 프로세서(110)는 텍스트 기반 오디오 생성 파라미터, 변환된 이미지 조건 형태 및 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출할 수 있다. 프로세서(110)는 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 이에, 프로세서(110)는 텍스트 및 이미지 기반 오디오 생성 파라미터를 통해 입력된 텍스트에 기초하여 오디오를 생성할 수 있다. 또한, 오디오 생성 모델을 트레이닝시키는 방식은 도 2에서 상세히 후술한다.
- [0055] 이를 통해 프로세서(110)는 오디오 생성 모델의 텍스트 기반 오디오 생성 파라미터를 우선적으로 결정하고, 결정된 텍스트 기반 오디오 생성 파라미터에 기초하여 텍스트 및 이미지 기반 오디오 생성 파라미터를 텍스트와 이미지를 이용한 트레이닝을 추가적으로 진행함에 따라 결정하므로, 전체적인 트레이닝 연산이 간단해질 수 있고 전체적인 트레이닝 과정 또한 빠르게 진행할 수 있다.

[0056] 도 2는 일 실시예에 따른 오디오 생성 모델의 아키텍처를 예시적으로 나타낸다.

[0057] 우선, 프로세서(110)는 텍스트 인코더(210)에 어느 트레이닝 데이터 세트의 텍스트 데이터(10)를 입력하여 텍스트 임베딩(embedding)을 추출할 수 있다.

[0058] 프로세서(110)는 어느 트레이닝 데이터 세트의 오디오 데이터를 참값(Ground Truth)으로 하여, 오디오 생성 모델의 인코더(231)를 통해 픽셀 공간(pixel space)(230)의 고 차원 참값(21)을 잠재 공간(latent space)(220)의 저 차원 참값으로 변환할 수 있다. 프로세서(110)는 변환된 저 차원 참값을 오디오 생성 모델을 통해 디퓨전(diffusion)하여 노이즈가 삽입된 저 차원 참값을 생성할 수 있다. 프로세서(110)는 텍스트 임베딩을 조건으로 오디오 생성 모델을 통해 노이즈가 삽입된 저 차원 참값을 디노이징(denoising)하여 노이즈가 제거된 저 차원 예측값을 생성할 수 있다. 프로세서(110)는 오디오 생성 모델의 디코더(232) 통해 저 차원 예측값을 고 차원 예측값(22)(Prediction)으로 변환할 수 있다. 프로세서(110)는 수학적식 1을 이용하여 고 차원 참값(21)과 고 차원 예측값(22) 간의 텍스트 기반 손실을 산출할 수 있다.

수학적식 1

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t, c} [\|\epsilon - \epsilon_{\theta}(\tilde{z}_t, t, c(d))\|_2^2]$$

[0059]

[0060] 수학적식 1에서, $\tilde{z}_t$ 는 노이즈가 추가된 저 차원 오디오 데이터를 나타내고, t 는 프로세스 시간을 나타내고, $c(d)$ 는 조건으로서, c 는 사전-트레이닝된 트레이닝 가능한 텍스트 인코더(210)를 나타내고, d 는 영상의 설명(description)인 텍스트 데이터를 나타낼 수 있다. 설명(description)은 이미지만을 설명하는 텍스트인 캡션(caption)과 달리 스토리를 포함하는 텍스트를 의미할 수 있다.

[0061] 프로세서(110)는 이미지 인코더(240)에 이미지 데이터(30)를 입력하여 이미지 임베딩을 추출할 수 있다. 이후, 프로세서(110)는 복제 모델(250)을 이용하여 텍스트 기반 손실을 최소화함에 따라 결정된 텍스트 기반 오디오 생성 파라미터를 동결(freezing)(예: 저장)하고 동결된 텍스트 기반 오디오 생성 파라미터에 기초하여 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 다시 말하자면, 텍스트 기반 오디오 생성 파라미터를 생성한 원 모델은 동결되어 더 이상 텍스트 기반 오디오 생성 파라미터를 업데이트하지 않도록 하고, 원 모델을 복사한 트레이닝 가능한 복사 모델을 이용하여 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 복제 모델(250)은 제1 제로 컨볼루션 레이어(Zero Convolution layer), 트레이닝 가능한 복사 모델 및 제2 제로 컨볼루션 레이어로 구성되고, 제로 컨볼루션 레이어는 트레이닝할수록 더욱 고품질의 결과를 생성할 수 있도록 한다. 이에 복사 모델은 제로 컨볼루션 레이어로 원 모델과 연결됨에 따라 프로세서(110)가 복사 모델을 트레이닝시키면 연결된 제로 컨볼루션 레이어 또한 트레이닝될 수 있다. 또한, 이는 텍스트 기반 오디오 생성 파라미터를 이용한 오디오 생성 구조가 유지되면서 이에 기초한 텍스트 및 이미지 기반 오디오 생성 파라미터를 이용한 오디오 생성 구조를 추가적으로 형성될 수 있음을 의미할 수 있다.

[0062] 또한, 프로세서(110)는 음악 생성 모델을 트레이닝하기 전에 사전-트레이닝된 텍스트 인코더(210) 및 사전-트레이닝된 이미지 인코더(240)를 트레이닝시킬 수 있다. 프로세서(110)는 텍스트 인코더(210) 및 이미지 인코더(240)를 음악 생성 모델과 동일한 트레이닝 데이터 세트를 이용할 수 있으나, 이에 한정하지 않고 별도의 트레이닝 데이터를 이용하여 트레이닝시킬 수 있다. 또한, 텍스트 인코더(210) 및 이미지 인코더(240)의 트레이닝은 프로세서(110)에서 수행되는 것으로 한정하지 않고, 별도의 프로세서를 통해 트레이닝된 후 획득한 파라미터 값을 장치(100)가 수신하여 프로세서(110)를 통해 텍스트 인코더(210) 및 이미지 인코더(240)에 적용할 수 있다. 또한, 텍스트 인코더(210) 및 이미지 인코더(240)는 텍스트에서 단어를 추출하거나 이미지에서 이미지 부분을 추출하고, 추출된 단어 및 이미지에서 숫자로 변환하며, 변환된 숫자를 잠재 벡터(latent vector)의 형태인 텍스트 임베딩으로 생성할 수 있다.

[0063] 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정하기 위하여, 프로세서(110)는 텍스트 기반 오디오 생성 파라미터를 결정하는 방식과 유사하게 진행할 수 있다. 프로세서(110)는 오디오 데이터를 참값(Ground Truth)으로 하여, 오디오 생성 모델의 인코더(231)를 통해 픽셀 공간(230)의 고 차원 참값(21)을 잠재 공간(220)의 저 차원 참값으로 변환할 수 있다. 프로세서(110)는 변환된 저 차원 참값을 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성할 수 있다. 프로세서(110)는 이미지 임베딩 및 텍스트 임베딩을 조건으로

로 텍스트 기반 오디오 생성 파라미터가 적용된 오디오 생성 모델을 통해 노이즈가 삽입된 저 차원 참값을 디노이징하여 노이즈가 제거된 저 차원 예측값을 생성할 수 있다. 프로세서(110)는 오디오 생성 모델의 디코더(232) 통해 저 차원 예측값을 고 차원 예측값(22)으로 변환할 수 있다. 프로세서(110)는 수학적 2를 이용하여 고 차원 참값(21)과 고 차원 예측값(22) 간의 텍스트 및 이미지 기반 손실을 산출할 수 있다.

수학적 2

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t, c, c_{img}} [\|\epsilon - \epsilon_{\theta}(\tilde{z}_t, t, c(d), c_{img}(f))\|_2^2]$$

수학적 2에서, $c_{img}(f)$ 는 수학적 1의 $c(d)$ 와 같이 조건으로서, c_{img} 는 사전-트레이닝된 트레이닝 가능한 이미지 인코더(240)를 나타내고, f 는 트레이닝 데이터 세트의 프레임 시퀀스(예: 이미지 데이터)를 나타낼 수 있다.

또한, 프로세서(110)는 추가로 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 프로세서(110)는 사전-트레이닝된 텍스트 인코더(210)를 통해 "Audio" 단어(31)에서 이미지 기반 텍스트 임베딩을 추출할 수 있다. "Audio" 단어(31)는 이미지 기반 오디오 생성 파라미터 결정 시 오디오에 대한 텍스트 정보를 제공하지 않도록 입력되는 조건일 수 있다. 이는, 예를 들어, "Audio" 단어가 아닌 텍스트 데이터(10)가 입력되면 이미지 기반 오디오 생성 파라미터가 아닌 텍스트 및 이미지 기반 오디오 생성 파라미터가 결정될 수 있기 때문이다.

프로세서(110)는 추출된 이미지 기반 텍스트 임베딩, 텍스트 기반 오디오 생성 파라미터 및 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 오디오 생성 모델의 이미지 기반 손실을 산출할 수 있다. 프로세서(110)는 산출된 이미지 기반 손실이 최소화되도록 오디오 생성 모델을 트레이닝시켜 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 프로세서(110)는 이미지 기반 오디오 생성 파라미터를 통해 입력된 이미지에 기초하여 오디오를 생성할 수 있다.

도 3은 일 실시예에 따른 오디오 생성 모델의 트레이닝을 위한 트레이닝 데이터 세트를 예시적으로 나타낸다.

프로세서(110)는 오디오 생성 모델을 트레이닝시키기 이전에, 트레이닝 데이터 세트를 생성할 수 있다. 프로세서(110)는 영상 데이터로부터 제1 지속 시간(예: 150초)의 복수의 제1 클립 데이터들(310)을 생성할 수 있다. 프로세서(110)는 영상 데이터로부터 제1 지속 시간보다 짧은 시간 길이인 제2 지속 시간(예: 3초 내지 12초)의 복수의 제2 클립 데이터들(320)을 생성할 수 있다. 예를 들어, 프로세서(110)는 복수의 제1 클립 데이터들(310)에 기초하여 22,330 개의 트레이닝 데이터 세트를 생성하고, 복수의 제2 클립 데이터들(320)에 기초하여 5,024 개의 트레이닝 데이터 세트를 생성할 수 있다. 프로세서(110)는 동일한 영상 데이터에 각기 다른 방식을 이용하여 복수의 제1 클립 데이터들(310)과 복수의 제2 클립 데이터들(320)을 생성하고 이를 병합시켜 트레이닝 데이터 세트를 생성하므로, 오디오 생성 모델이 상호 보완적으로 트레이닝될 수 있다. 이는 제1 클립 데이터의 텍스트(313)와 제2 클립 데이터의 텍스트(323)가 다른 표현 형식을 가지고 있어(예: 텍스트(313)는 서술형, 텍스트(323)은 단문형), 사용자가 다양한 형태의 스토리를 입력하여도 이에 기초로 적절한 오디오를 생성할 수 있음을 의미한다.

프로세서(110)는 제1 지속 시간에 대응되는 시간 길이가 되도록 복수의 제2 클립 데이터들(320)을 연결시킬 수 있다. 예를 들어, 복수의 제1 클립 데이터들(310)의 지속시간이 150초이고, 복수의 제2 클립 데이터들(320)의 지속시간이 3초인 경우, 프로세서(110)는 50개의 제2 클립 데이터들(320)을 연결시킬 수 있다.

프로세서(110)는 복수의 제1 클립 데이터들과 연결된 복수의 제2 클립 데이터들을 병합시켜 복수의 원시 클립 데이터를 생성할 수 있다. 예를 들어, 3,807 개의 영상(예: 영화)에 기초하여 생성된 복수의 원시 클립 데이터는 27,354 개일 수 있다.

프로세서(110)는 복수의 원시 클립 데이터로부터 추출된 이미지 데이터(311, 321) 및 음성 분리 오디오 데이터(312, 322)와 이에 기반한 텍스트 데이터(313, 323)에 기초하여 복수의 트레이닝 데이터 세트를 생성할 수 있다. 트레이닝 데이터 세트는 오디오 데이터, 이미지 데이터 및 텍스트 데이터 쌍(pair)으로 형성될 수 있다. 예를 들어, 각 원시 클립 데이터의 텍스트는 평균 88단어로 구성되고, 복수의 제1 클립 데이터들(310)의 평균 1,378.7 단어로 구성되며, 연결되기 전 복수의 제2 클립 데이터들(320)은 평균 83.2 단어로 구성될 수 있다.

또한, 트레이닝 데이터 세트의 텍스트 데이터(예: 스토리)는 27,354개이고, 오디오 데이터(312, 322)는 평균 약 150초로 구성되어 총 누적 시간이 984 시간이며, 하나의 텍스트 데이터(313, 323)를 기준으로 19.6개의 이미지 데이터(311, 321)를 가질 수 있다.

[0073] 또한, 프로세서(110)는 음성 분리 오디오 데이터를 생성한 이후, 음성 분리 오디오 데이터를 스펙트로그램(Spectrogram) 이미지로 변환할 수 있다. 음성 분리 오디오 데이터에 대해서는 도 4에서 후술한다. 스펙트로그램은 오디오 분석을 위해 전처리되는 형태로서, 시간에 따른 주파수의 변화를 그래픽으로 나타낼 수 있다. 오디오 정보는 X축을 시간, Y축을 주파수(frequency), Z축을 진폭(Amplitude)으로 표시될 수 있다. 이에 스펙트로그램은 오디오 신호가 시각적으로 표현되어 주파수 내용 및 강도를 시간에 따라 확인할 수 있으므로, 이에 기초로 사운드의 주파수 성분을 분석할 수 있다.

[0074] 도 4는 일 실시예에 따른 오디오 생성 모델의 트레이닝을 위한 트레이닝 데이터 세트의 오디오 데이터, 이미지 데이터 및 텍스트 데이터가 생성되는 과정을 예시적으로 나타낸다.

[0075] 프로세서(110)는 원시 클립 데이터로부터 비디오 프레임들(411)을 추출하고, 추출된 비디오 프레임들(411)을 클러스터링(412)하여 하나 이상의 이미지 데이터(413)를 추출할 수 있다. 영상에는 중복되거나 유사한 이미지가 많으므로 스토리를 대표하는 핵심 이미지만 추출하는 것이 필요할 수 있다. 이에 프로세서(110)는 클러스터링을 수행할 수 있다. 프로세서(110)는 비디오 프레임들의 이미지 핑거프린트(fingerprint)의 유사도에 기초하여 이미지를 제거하고 남은 이미지를 클러스터링하여 이미지 데이터(413)로 구성할 수 있다. 예를 들어, 프로세서(110)는 모든 이미지의 핑거프린트에서 유사도가 0.5 이상인 이미지를 제거할 수 있다. 이를 통해 정적인 장면과 동적인 장면에 있어서, 정적인 장면이 동적인 장면보다 더 적은 개수의 이미지 데이터가 구성될 수 있다. 이는 최소한의 이미지 데이터를 이용하여 스토리 라인을 강조할 수 있음을 의미한다.

[0076] 또한, 스토리가 포함된 오디오는 특성 상 단일 장면이 지나갈 때까지 일관성을 거의 유지하지 않아 오디오의 길이를 분할(slice)하는 것이 필요할 수 있다. 예를 들어, 프로세서(110)는 오디오 데이터(312, 322)의 길이를 15초로 분할할 수 있다. 이미지 데이터(311, 321)는 각 이미지 내 의미 변화를 부분적으로 반영하므로 추출된 두 이미지 간의 오디오 데이터를 기본 단위로 형성할 수 있다. 이에 프로세서(110)는 단위 오디오가 15초가 되도록 누적 및/또는 연결할 수 있다. 만약, 오디오 자체가 15초보다 긴 경우, 해당 부분은 중요한 시각적 변화가 없어 길고 일관된 오디오일 가능성이 높으므로 분할하지 않을 수 있다. 또한, 15초의 오디오에 약간(예: 78Hz 미만의)의 오디오가 포함된 경우, 해당 세그먼트(segment)는 트레이닝에 적절하지 않으므로 제거될 수 있다. 15초 길이의 무작위로 분할된 오디오 데이터(312, 322)는 총 185,341개로, 184,462개는 트레이닝에 사용되고, 879개는 테스트에 사용될 수 있다. 이에 오디오 데이터(312, 322)의 길이 분할은 길이가 긴 오디오를 사용하는 것보다 하드웨어의 용량 부하가 감소하며 이를 처리하는 속도 또한 빨라질 수 있다.

[0077] 다시 말하자면, 프로세서(110)는 클러스터링을 통해 영상을 분할할 수 있다. 예를 들어, 180초 길이의 영화(예: 원시 클립 데이터)가 있는 경우, 프로세서(110)는 클러스터링을 통해 다섯개(예: 00:00, 00:10, 00:13, 00:50, 03:00) 시간에 관한 이미지를 이미지 데이터(413)로 구성할 수 있다. 이후, 프로세서(110)는 네개(예: 00:00~00:10, 00:10~00:13, 00:13~00:50, 00:50~03:00)의 시간에 대하여 원시 클립 데이터를 나눌 수 있다. 프로세서(110)는 원시 클립 데이터에서 각각 10초, 3초, 37초, 130초 길이의 음원을 추출할 수 있다. 프로세서(110)는 00:00~00:10, 00:10~00:13이 각각 10초, 3초 길이를 가지므로 15초에 대응하기 위해 연결하는 것이 필요할 수 있다. 00:13~00:50, 00:50~03:00은 각각 37초, 130초 길이이고 이는 이미지 변화가 거의 없는 영상에 해당될 수 있기 때문에 추가적인 작업을 필요로 하지 않는다. 이에 이러한 예외를 제외하고는 전반적으로 하나의 영화(예: 원시 클립 데이터)에서 15초 내외의 음원들이 생성될 수 있다.

[0078] 프로세서(110)는 원시 클립 데이터로부터 원시 오디오(421)를 추출하고, 추출된 원시 오디오(421)에서 음성(speech)(422)을 분리(decoupling)하여 음성 분리 오디오 데이터(423)를 생성할 수 있다. 이는 오디오 생성 모델이 스토리 라인에 기초하여 배경음(Background sound)을 생성해야 하므로, 스토리 라인이 포함된 음성을 트레이닝 데이터에 제외시키는 것을 의미할 수 있다. 음성 분리 오디오 데이터(423)(예: 배경음)은 다양한 일상의 소리와 배경 음악이 섞여 있을 수 있으나, 이러한 특징은 오디오를 이용한 스토리 이해에 더 효과적일 수 있다.

[0079] 프로세서(110)는 이미지 데이터와 음성 분리 오디오 데이터에 복수의 제1 클립 데이터들 및 복수의 제2 클립 데이터들 생성 시 제공되는 텍스트 데이터를 대응시켜 오디오 데이터, 이미지 데이터 및 텍스트 데이터의 쌍의 트레이닝 데이터 세트를 생성할 수 있다. 예를 들어, 프로세서(110)는 트레이닝 데이터를 생성함에 있어서 우선위에서 언급한 바와 같이 이미지 데이터를 기준으로 오디오 데이터가 슬라이스된 후 다시 병합할 수 있다. 프로세서(110)는 00:00~00:10, 00:10~00:13이 합산된 00:00~00:13음원에 기초하여 이미지 시퀀스를 생성할 수 있

다. 프로세서(110)는 00:00, 00:10 시점의 이미지를 연결하여 이미지 데이터(예: 시퀀스)를 생성할 수 있다. 프로세서(110)는 00:13~00:50 음원에 기초하여 00:13 시점의 이미지를 이미지 데이터로 할 수 있다. 이후, 프로세서(110)는 병합된 이미지 데이터와 오디오 데이터에 텍스트 데이터를 대응시킬 수 있다. 제1 클립 데이터들의 텍스트 데이터는 제2 클립 데이터들의 텍스트 데이터보다 더 많은 텍스트 양을 가지고 있을 수 있다. 제1 클립 데이터들의 텍스트 데이터의 경우, 프로세서(110)는 음원 개수에 따라 기계적으로 문장 또는 단어를 나누어 각각의 시간 단위에 따라 텍스트 데이터로 할 수 있다. 또한, 제2 클립 데이터들의 텍스트 데이터의 경우, 프로세서(110)는 텍스트 양이 부족하므로, 단어를 나누거나 전체 텍스트 데이터를 동일하게 각각의 시간 단위에 따라 텍스트 데이터로 할 수 있다. 예를 들어, 180초 길이의 제1 클립 데이터들인 경우, "풋볼 코치와 보조원, 엘라배마 선수들이 그를 응원합니다. 그는 경기장을 가로질러 달려갑니다. 그는 수비 선수들을 빠르게 지나쳐 나갑니다." 인 텍스트 데이터가 00:00~00:13, 00:13~00:50, 00:50~03:00 음성에 각각 "풋볼 코치와 보조원, 엘라배마 선수들이 그를 응원합니다.", "그는 경기장을 가로질러 달려갑니다.", "그는 수비 선수들을 빠르게 지나쳐 나갑니다." 로 대응될 수 있다. 이와 다르게 제2 클립 데이터들인 경우, "그는 응원을 받으며 경기장을 달려갑니다."인 텍스트 데이터가 00:00~00:13, 00:13~00:50, 00:50~03:00 음성에 각각 "그는", "응원을 받으며", "경기장을 달려나갑니다." 또는 셋 다 그는 응원을 받으며 경기장을 달려갑니다."로 대응될 수 있다. 프로세서(110)는 이러한 트레이닝 데이터 세트를 이용하여 오디오 생성 모델을 트레이닝시켜 스토리에 기반한 배경음을 생성할 수 있다.

[0080] 다양한 실시예에 따르면, 프로세서(110)는 메모리(120)에 저장된 컴퓨터로 읽을 수 있는 코드(예를 들어, 소프트웨어) 및 프로세서(110)에 의해 유발된 인스트럭션(instruction)들을 실행할 수 있다. 프로세서(110)는 목적하는 동작들(desired operations)을 실행시키기 위한 물리적인 구조를 갖는 회로를 가지는 하드웨어로 구현된 데이터 처리장치일 수 있다. 목적하는 동작들은 예를 들어, 프로그램에 포함된 코드(code) 또는 인스트럭션들(instructions)을 포함할 수 있다. 하드웨어로 구현된 데이터 처리 장치는 예를 들어, 마이크로프로세서(microprocessor), 중앙 처리 장치(central processing unit), 프로세서 코어(processor core), 멀티-코어 프로세서(multi-core processor), 멀티프로세서(multiprocessor), ASIC(Application-Specific Integrated Circuit), FPGA(Field Programmable Gate Array)를 포함할 수 있다.

[0081] 메모리(120)는 오디오 생성 모델, 오디오 생성 모델의 트레이닝에 필요한 트레이닝 데이터 세트 및 오디오 생성 모델의 트레이닝에 의해 생성된 파라미터들이 저장될 수 있다.

[0082] 또한, 메모리(120)는 프로세서(110)에 의해 실행가능한 인스트럭션들(예: 프로그램)을 저장할 수 있다. 예를 들어, 인스트럭션들은 프로세서(110)의 동작 및/또는 프로세서(110)의 각 구성의 동작을 실행하기 위한 인스트럭션들을 포함할 수 있다.

[0083] 다양한 실시예에 따르면, 메모리(120)는 휘발성 메모리 장치 또는 불휘발성 메모리 장치로 구현될 수 있다. 휘발성 메모리 장치는 DRAM(dynamic random access memory), SRAM(static random access memory), T-RAM(thyristor RAM), Z-RAM(zero capacitor RAM), 또는 TTRAM(Twin Transistor RAM)으로 구현될 수 있다. 불휘발성 메모리 장치는 EEPROM(Electrically Erasable Programmable Read-Only Memory), 플래시(flash) 메모리, MRAM(Magnetic RAM), 스핀전달토크 MRAM(Spin-Transfer Torque(STT)-MRAM), Conductive Bridging RAM(CBRAM), FeRAM(Ferroelectric RAM), PRAM(Phase change RAM), 저항 메모리(Resistive RAM(RRAM)), 나노 튜브 RRAM(Nanotube RRAM), 폴리머 RAM(Polymer RAM(PoRAM)), 나노 부유 게이트 메모리(Nano Floating Gate Memory(NFGM)), 홀로그래픽 메모리(holographic memory), 분자 전자 메모리 소자(Molecular Electronic Memory Device), 및/또는 절연 저항 변화 메모리(Insulator Resistance Change Memory)로 구현될 수 있다.

[0084] 도 5는 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법의 흐름도를 도시한다.

[0085] 단계(510)에서, 프로세서는 사전-트레이닝된(pre-trained) 텍스트 인코더를 통해 오디오 데이터, 이미지 데이터 및 텍스트 데이터를 포함하는 트레이닝 데이터 세트 중 텍스트 데이터를 오디오 생성 모델의 텍스트 조건 형태로 변환할 수 있다. 프로세서는 텍스트 인코더에 텍스트 데이터를 입력하여 텍스트 임베딩(embedding)을 추출할 수 있다.

[0086] 단계(520)에서, 프로세서는 변환된 텍스트 조건 형태 및 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 오디오 생성 모델의 텍스트 기반 손실(loss)을 산출할 수 있다.

[0087] 단계(530)에서, 프로세서는 산출된 텍스트 기반 손실이 최소화되도록 오디오 생성 모델을 트레이닝시켜 텍스트 기반 오디오 생성 파라미터(parameter)를 결정할 수 있다.

- [0088] 단계(540)에서, 프로세서는 트레이닝 데이터 세트 중 이미지 데이터를 사전-트레이닝된 이미지 인코더를 통해 오디오 생성 모델의 이미지 조건 형태로 변환할 수 있다. 프로세서는 이미지 인코더에 이미지 데이터를 입력하여 이미지 임베딩을 추출할 수 있다.
- [0089] 단계(550)에서, 프로세서는 텍스트 기반 오디오 생성 파라미터, 변환된 이미지 조건 형태 및 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 오디오 생성 모델의 텍스트 및 이미지 기반 손실을 산출할 수 있다.
- [0090] 단계(560)에서, 프로세서는 산출된 텍스트 및 이미지 기반 손실이 최소화되도록 오디오 생성 모델을 트레이닝시켜 텍스트 및 이미지 기반 오디오 생성 파라미터를 결정할 수 있다.
- [0091] 프로세서는 추가로 이미지 기반 오디오 생성 파라미터를 결정할 수 있다. 프로세서는 텍스트 기반 오디오 생성 파라미터를 결정한 이후, 트레이닝 데이터 세트 중 이미지 데이터로부터 하나 이상의 단어들을 추출할 수 있다. 프로세서는 사전-트레이닝된 텍스트 인코더를 통해 추출된 하나 이상의 단어들에서 이미지 기반 텍스트 임베딩을 추출할 수 있다. 프로세서는 추출된 이미지 기반 텍스트 임베딩, 텍스트 기반 오디오 생성 파라미터 및 트레이닝 데이터 세트 중 오디오 데이터에 기초하여, 오디오 생성 모델의 이미지 기반 손실을 산출할 수 있다. 프로세서는 산출된 이미지 기반 손실이 최소화되도록 오디오 생성 모델을 트레이닝시켜 이미지 기반 오디오 생성 파라미터를 결정할 수 있다.
- [0092] 텍스트 기반 손실을 산출함에 있어서, 프로세서는 오디오 데이터를 참값(Ground Truth)으로 하여, 오디오 생성 모델의 인코더를 통해 픽셀 공간(pixel space)의 고 차원 참값을 잠재 공간(latent space)의 저 차원 참값으로 변환할 수 있다. 프로세서는 변환된 저 차원 참값을 오디오 생성 모델을 통해 디퓨전(diffusion)하여 노이즈가 삽입된 저 차원 참값을 생성할 수 있다. 프로세서는 텍스트 임베딩을 조건으로 오디오 생성 모델을 통해 노이즈가 삽입된 저 차원 참값을 디노이징(denoising)하여 노이즈가 제거된 저 차원 예측값을 생성할 수 있다. 프로세서는 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환할 수 있다. 프로세서는 고 차원 참값과 고 차원 예측값 간의 텍스트 기반 손실을 산출할 수 있다.
- [0093] 텍스트 및 이미지 기반 손실을 산출함에 있어서, 프로세서는 오디오 데이터를 참값으로 하여, 오디오 생성 모델의 인코더를 통해 픽셀 공간의 고 차원 참값을 잠재 공간의 저 차원 참값으로 변환할 수 있다. 프로세서는 변환된 저 차원 참값을 오디오 생성 모델을 통해 디퓨전하여 노이즈가 삽입된 저 차원 참값을 생성할 수 있다. 프로세서는 이미지 임베딩 및 텍스트 임베딩을 조건으로 텍스트 기반 오디오 생성 파라미터가 적용된 오디오 생성 모델을 통해 노이즈가 삽입된 저 차원 참값을 디노이징하여 노이즈가 제거된 저 차원 예측값을 생성할 수 있다. 프로세서는 오디오 생성 모델의 디코더를 통해 저 차원 예측값을 고 차원 예측값으로 변환할 수 있다. 프로세서는 고 차원 참값과 고 차원 예측값 간의 텍스트 및 이미지 기반 손실을 산출할 수 있다.
- [0094] 또한, 프로세서는 오디오 생성 모델을 트레이닝시키기 이전에, 트레이닝 데이터 세트를 생성할 수 있다. 트레이닝 데이터 세트를 생성함에 있어서, 프로세서는 영상 데이터로부터 제1 지속 시간의 복수의 제1 클립 데이터들을 생성할 수 있다. 또한, 프로세서는 영상 데이터로부터 제1 지속 시간보다 짧은 시간 길이인 제2 지속 시간의 복수의 제2 클립 데이터들을 생성할 수 있다. 그런 후, 프로세서는 제1 지속 시간에 대응되는 시간 길이가 되도록 복수의 제2 클립 데이터들을 연결시킬 수 있다. 프로세서는 복수의 제1 클립 데이터와 연결된 복수의 제2 클립 데이터들을 병합시켜 복수의 원시 클립 데이터를 생성할 수 있다. 이에 프로세서는 복수의 원시 클립 데이터로부터 추출된 이미지 데이터 및 음성 분리 오디오 데이터와 이에 기반한 텍스트 데이터에 기초하여 복수의 트레이닝 데이터 세트를 생성할 수 있다.
- [0095] 복수의 트레이닝 데이터 세트에 포함되는 이미지 데이터를 생성함에 있어서, 프로세서는 원시 클립 데이터로부터 비디오 프레임들을 추출하고, 추출된 비디오 프레임들을 클러스터링하여 하나 이상의 이미지 데이터를 추출할 수 있다. 예를 들어, 프로세서는 비디오 프레임들의 이미지 핑거프린트(fingerprint)의 유사도에 기초하여 이미지를 제거하고 남은 이미지를 클러스터링하여 이미지 데이터로 구성할 수 있다.
- [0096] 복수의 트레이닝 데이터 세트에 포함되는 오디오 데이터를 생성함에 있어서, 프로세서는 원시 클립 데이터로부터 원시 오디오를 추출하고, 추출된 원시 오디오에서 음성(speech)을 분리(decoupling)하여 음성 분리 오디오 데이터를 생성할 수 있다. 음성 분리 오디오 데이터를 생성한 이후, 프로세서는 음성 분리 오디오 데이터를 스펙트로그램(Spectrogram) 이미지로 변환할 수 있다.
- [0097] 복수의 트레이닝 데이터 세트를 생성함에 있어서, 프로세서는 이미지 데이터와 음성 분리 오디오 데이터에 복수의 제1 클립 데이터들 및 복수의 제2 클립 데이터들 생성 시 제공되는 텍스트 데이터를 대응시켜 오디오 데이터, 이미지 데이터 및 텍스트 데이터의 쌍의 트레이닝 데이터 세트를 생성할 수 있다.

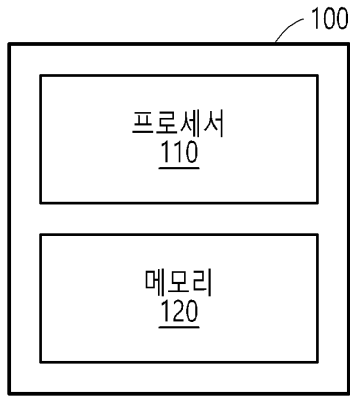
- [0098] 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법으로 트레이닝된 오디오 생성 모델은 이미지 시퀀스(예: 이미지 데이터) 및/또는 텍스트(예: 텍스트 데이터)의 스토리 컨텍스트가 주어지면 적절한 배경 오디오를 생성할 수 있다.
- [0099] 도 6은 일 실시예에 따른 오디오 생성 모델을 트레이닝시키는 방법으로 트레이닝된 오디오 생성 모델과 다른 음악 생성 모델의 FAD(Frechet Audio Distance)기반 평가 결과를 나타낸다.
- [0100] FAD(Frechet Audio Distance)는 모델의 음악 출력만 사용하여 계산할 수 있고, 임베딩의 평균과 공분산을 사용하여 계산되는 참조 없는 평가 지표로서, 값이 낮을수록 높은 성능을 나타낼 수 있다. 비교 모델인 Riffusion은 전체 모델 중 가장 높은 값인 20.114로 가장 낮은 성능을 가진 것으로 볼 수 있다. 그 다음, (620)은 이미지 데이터만 입력되어 배경음이 생성되는 모델로서, 전체 모델 중 두번째로 낮은 성능을 가질 수 있다. 또한, 또 다른 비교 모델인 MUSICGEN은 비록 (620)보다 높은 성능을 보여주고 있지만 이미지 및/또는 텍스트의 스토리(예: 맥락)에 적절한 효과음을 포함하는 배경음이 아닌 단순히 입력된 텍스트에 기초하여 음악을 생성함에 따라 생성되는 오디오가 다른 점에서 차이가 있다. (610)은 텍스트 데이터만 입력되어 배경음이 생성되는 모델로서, 전체 모델 중 두번째로 높은 성능을 가질 수 있다. 또한, (630)은 텍스트 데이터와 이미지 데이터가 입력되어 배경음이 생성되는 모델로서, FAD 값이 9.099로 전체 모델 중 가장 높은 성능을 가지고 있음을 확인할 수 있다.
- [0101] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 컨트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0102] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 저장될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0103] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [0104] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0105] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수

있다.

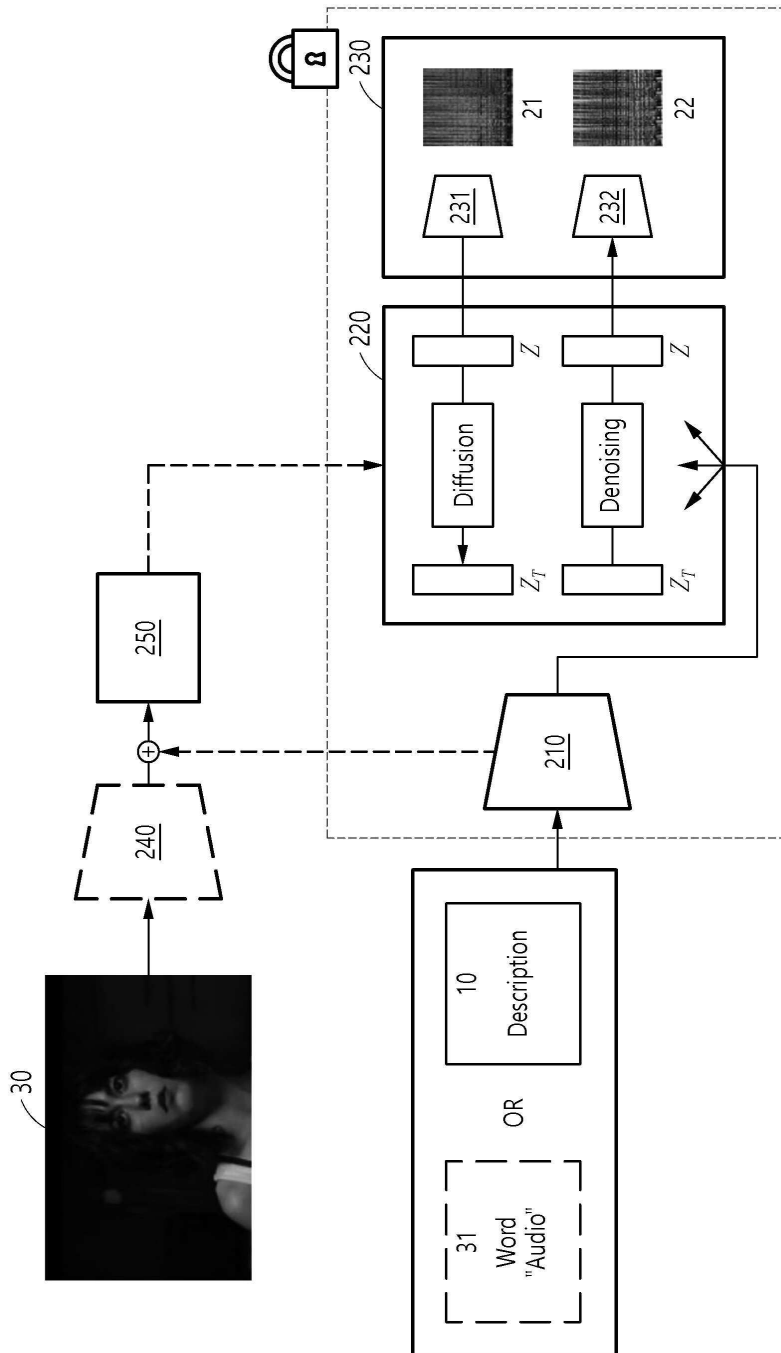
[0106] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

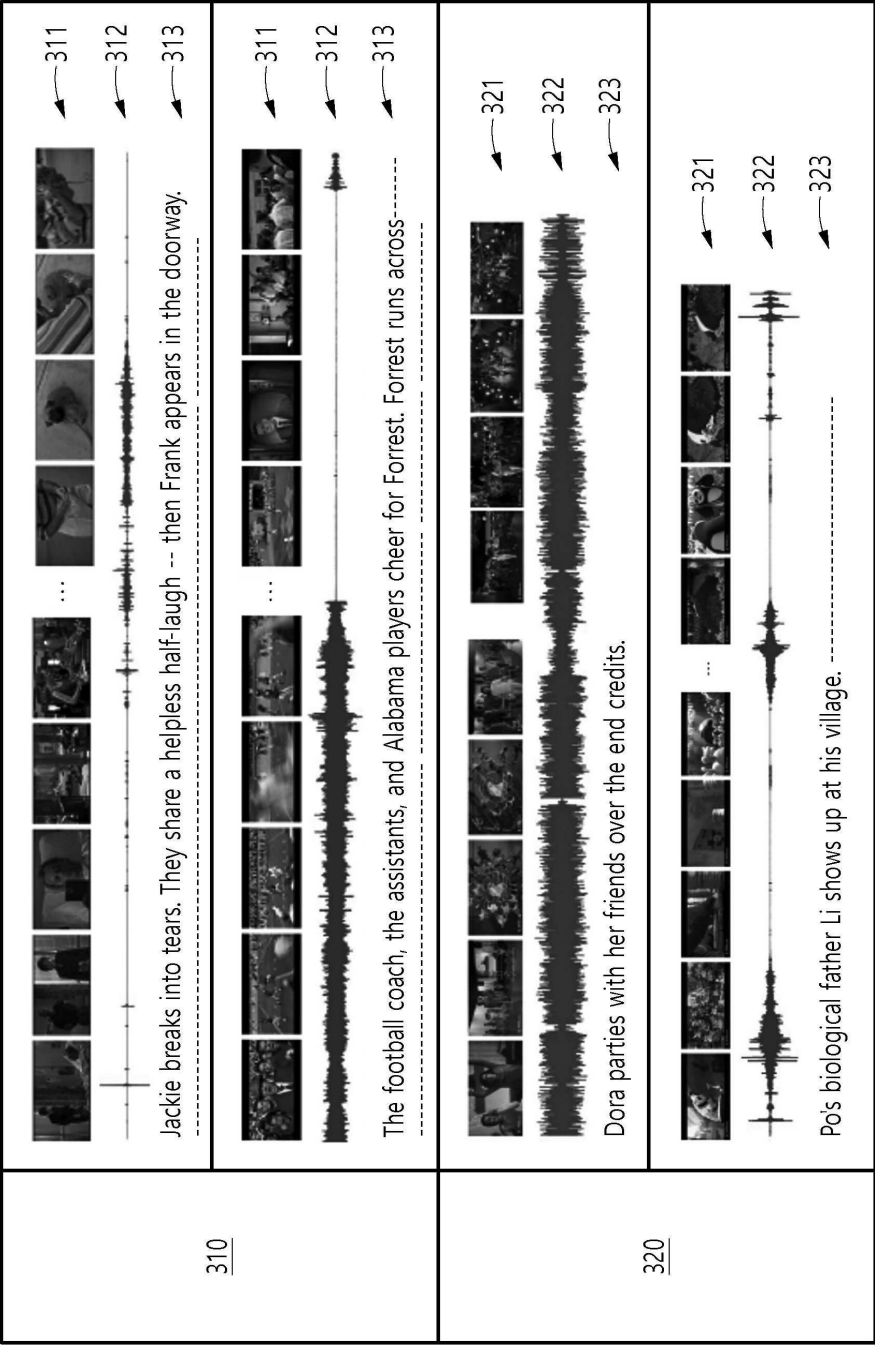
도면1



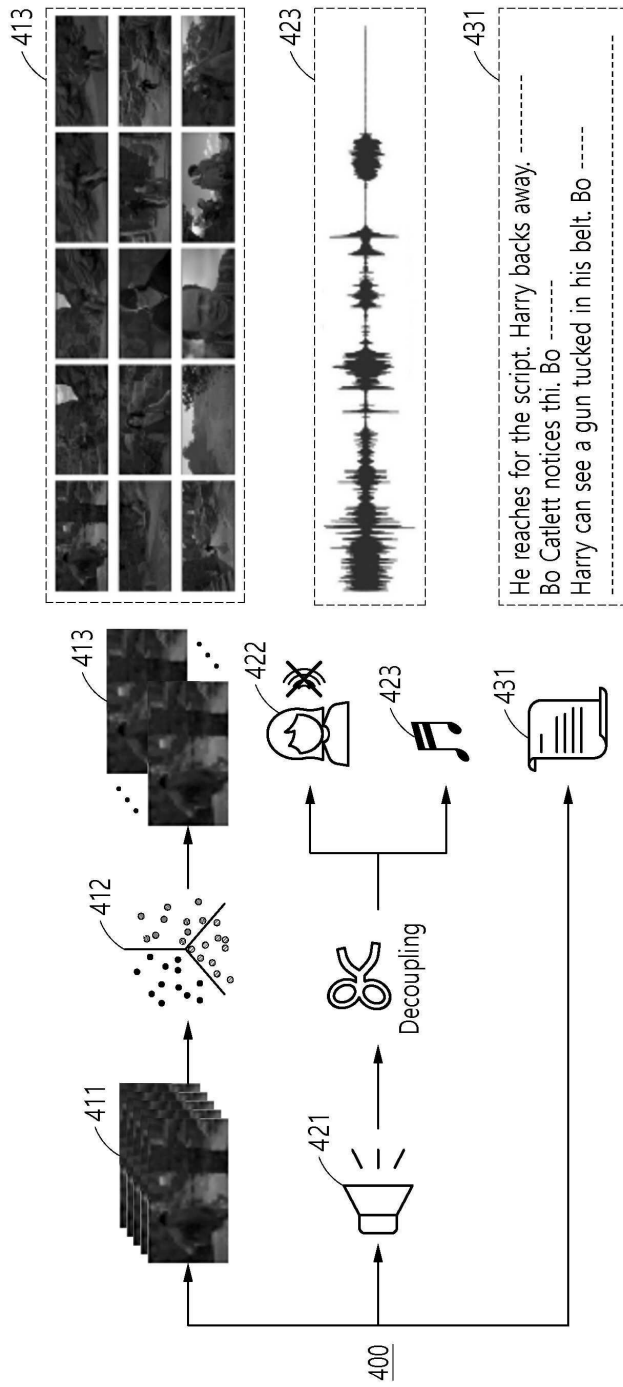
도면2



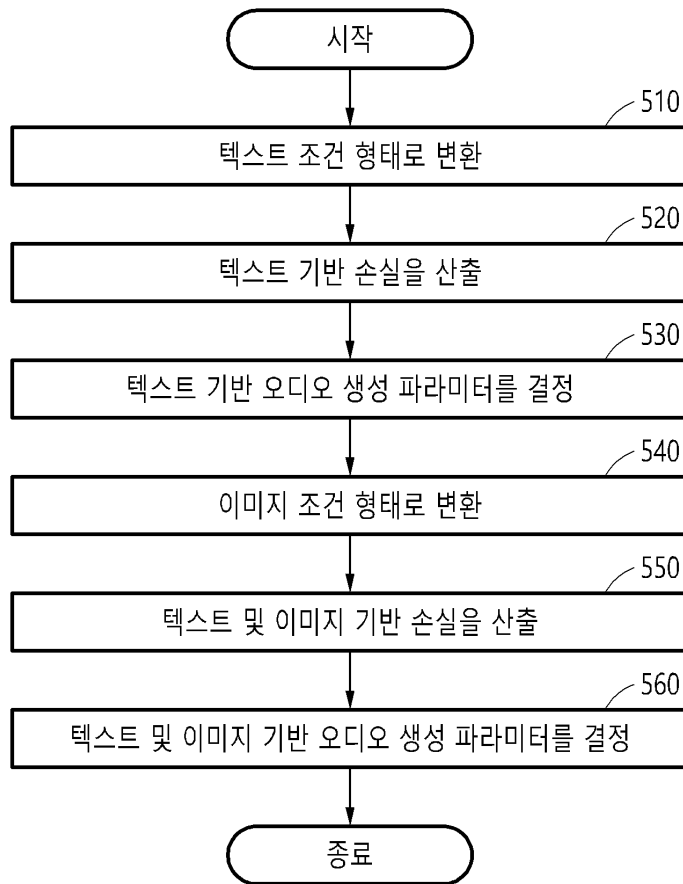
도면3



도면4



도면5



도면6

Model	$FAD_{VGGish} \downarrow$
Riffusion	20.114
MUSICGEN	11.680
<u>610</u>	10.935
<u>620</u>	18.324
<u>630</u>	9.099