

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0030359
(43) 공개일자 2025년03월05일

- (51) 국제특허분류(Int. Cl.)
G10L 21/038 (2013.01) *G06N 3/0455* (2023.01)
G10L 19/02 (2006.01) *G10L 25/18* (2013.01)
G10L 25/30 (2013.01)
- (52) CPC특허분류
G10L 21/038 (2013.01)
G06N 3/0455 (2023.01)
- (21) 출원번호 10-2023-0138366
(22) 출원일자 2023년10월17일
심사청구일자 2023년10월17일
(30) 우선권주장
1020230111458 2023년08월24일 대한민국(KR)
- (71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
- (72) 발명자
이성환
서울특별시 강남구 언주로30길 13, B-2506(도곡동, 대림아크로빌, 대림아크로텔)
김승빈
경기도 고양시 일산동구 위시티1로 7,503동 903호(식사동, 위시티블루밍5단지)
(뒷면에 계속)
- (74) 대리인
윤귀상

전체 청구항 수 : 총 14 항

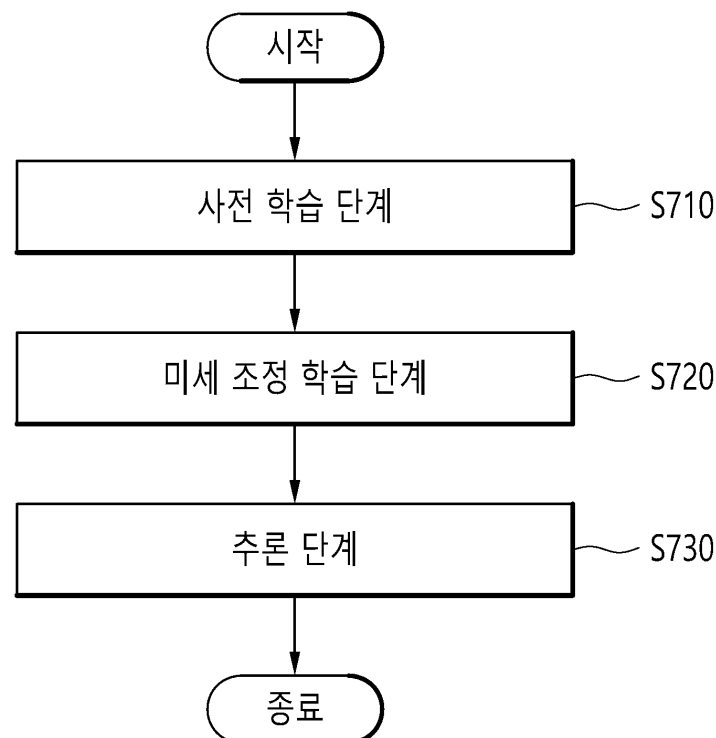
(54) 발명의 명칭 마스크된 오토인코더에 기초하여 고주파수 대역폭을 확장하는 오디오 초해상도 방법 및 장치

(57) 요약

고주파수 영역의 오디오 신호를 복원하는 방법 및 장치가 개시된다. 본 개시의 일 실시예에 따른, 장치에 의해 수행되는, 고주파수 영역의 오디오 신호를 복원하는 방법은, 제1 오디오 신호 학습 데이터에 대해 무작위 마스킹(masking)을 수행한 제1 마스킹 데이터를 제1 인코더(encoder)에 입력하여 제1 인코딩 데이터를 획득하고, 상기

(뒷면에 계속)

대표도 - 도7



제1 인코딩 데이터를 디코딩한 데이터 및 타겟 데이터 간의 차이에 기초하여 상기 제1 인코더를 학습시키는 사전 학습 단계; 제2 오디오 신호 학습 데이터에 대해 마스킹을 수행한 제2 마스킹 데이터를 제2 인코더에 입력하여 제2 인코딩 데이터를 획득하고, 상기 제2 인코딩 데이터를 오디오 신호 생성부에 입력하여 출력한 오디오 신호 데이터 및 상기 타겟 데이터에 기초하여 상기 제2 인코더 및 상기 오디오 신호 생성부를 학습시키는 미세 조정 학습 단계; 및 제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 상기 학습된 제2 인코더에 입력하여 제3 인코딩 데이터를 획득하고, 상기 제3 인코딩 데이터를 통해 상기 제1 주파수 해상도보다 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득하는 추론 단계를 포함하고, 상기 사전 학습 단계를 통해 학습된 상기 제1 인코더의 가중치는 상기 제2 인코더로 전이될 수 있다.

(52) CPC특허분류

G10L 19/0204 (2013.01)

G10L 25/18 (2013.01)

G10L 25/30 (2013.01)

(72) 발명자

이상훈

서울특별시 성북구 고려대로27길 34-8 (안암동5가)

최하영

경상북도 포항시 북구 흥해읍 한동로 558, 508호

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
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193973
과제번호	2019-0-00079-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(고려대학교)
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

장치에 의해 수행되는, 고주파수 영역의 오디오 신호를 복원하는 방법에 있어서, 상기 방법은:

제1 오디오 신호 학습 데이터에 대해 무작위 마스킹(masking)을 수행한 제1 마스킹 데이터를 제1 인코더(encoder)에 입력하여 제1 인코딩 데이터를 획득하고, 상기 제1 인코딩 데이터를 디코딩한 데이터 및 타겟 데이터 간의 차이에 기초하여 상기 제1 인코더를 학습시키는 사전 학습 단계;

제2 오디오 신호 학습 데이터에 대해 마스킹을 수행한 제2 마스킹 데이터를 제2 인코더에 입력하여 제2 인코딩 데이터를 획득하고, 상기 제2 인코딩 데이터를 오디오 신호 생성부에 입력하여 출력한 오디오 신호 데이터 및 상기 타겟 데이터에 기초하여 상기 제2 인코더 및 상기 오디오 신호 생성부를 학습시키는 미세 조정 학습 단계; 및

제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 상기 학습된 제2 인코더에 입력하여 제3 인코딩 데이터를 획득하고, 상기 제3 인코딩 데이터를 통해 상기 제1 주파수 해상도보다 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득하는 추론 단계를 포함하고,

상기 사전 학습 단계를 통해 학습된 상기 제1 인코더의 가중치는 상기 제2 인코더로 전이되는, 방법.

청구항 2

제1항에 있어서,

상기 제1 오디오 신호 학습 데이터, 상기 제2 오디오 신호 학습 데이터, 및 상기 오디오 신호 입력 데이터 각각은, 복수의 패치로 분할된 멜-스펙트로그램 형태를 가지는, 방법.

청구항 3

제2항에 있어서,

제1 오디오 신호 학습 데이터에 대해 무작위 마스킹을 수행하는 단계는,

상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치로부터 무작위로 적어도 하나의 패치를 선택하는 단계; 및

상기 적어도 하나의 패치에 대해 마스킹을 수행하는 단계를 포함하는, 방법.

청구항 4

제1항에 있어서,

상기 제2 오디오 신호 학습 데이터에 대해 무작위 마스킹 또는 특정 밴드 마스킹 중의 적어도 하나가 수행되는, 방법.

청구항 5

제4항에 있어서,

상기 제2 오디오 신호 학습 데이터에 대해 특정 밴드 마스킹을 수행하는 단계는,

상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치 중 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치를 식별하는 단계;

상기 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치에 대해 마스킹을 수행하는 단계를 포함하는, 방법.

청구항 6

제5항에 있어서,

상기 오디오 신호 입력 데이터와 연관된 샘플링 속도 유형의 수에 기초하여 상기 복수의 패치 중 마스킹되는 패치의 비율 패턴의 수가 결정되는, 방법.

청구항 7

제1항에 있어서,

상기 제1 인코더 및 상기 제2 인코더는, 비전 트랜스포머(vision transformer) 구조를 포함하는, 방법.

청구항 8

고주파수 영역의 오디오 신호를 복원하는 장치에 있어서, 상기 장치는:

제1 오디오 신호 학습 데이터에 대해 무작위 마스킹(masking)을 수행한 제1 마스킹 데이터를 제1 인코더(encoder)에 입력하여 제1 인코딩 데이터를 획득하고, 상기 제1 인코딩 데이터를 디코딩한 데이터 및 타겟 데이터 간의 차이에 기초하여 상기 제1 인코더를 학습시키는 사전 학습부;

제2 오디오 신호 학습 데이터에 대해 마스킹을 수행한 제2 마스킹 데이터를 제2 인코더에 입력하여 제2 인코딩 데이터를 획득하고, 상기 제2 인코딩 데이터를 오디오 신호 생성부에 입력하여 출력한 오디오 신호 데이터 및 상기 타겟 데이터에 기초하여 상기 제2 인코더 및 상기 오디오 신호 생성부를 학습시키는 미세 조정 학습부; 및

제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 상기 학습된 제2 인코더에 입력하여 제3 인코딩 데이터를 획득하고, 상기 제3 인코딩 데이터를 통해 상기 제1 주파수 해상도보다 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득하는 추론부를 포함하고,

상기 사전 학습 단계를 통해 학습된 상기 제1 인코더의 가중치는 상기 제2 인코더로 전이되는, 장치.

청구항 9

제8항에 있어서,

상기 제1 오디오 신호 학습 데이터, 상기 제2 오디오 신호 학습 데이터, 및 상기 오디오 신호 입력 데이터 각각은, 복수의 패치로 분할된 멜-스펙트로그램 형태를 가지는, 장치.

청구항 10

제9항에 있어서,

상기 사전 학습부에 포함된 마스킹부는:

상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치로부터 무작위로 적어도 하나의 패치를 선택하고; 및

상기 적어도 하나의 패치에 대해 마스킹을 수행하는, 장치.

청구항 11

제8항에 있어서,

상기 제2 오디오 신호 학습 데이터에 대해 무작위 마스킹 또는 특정 밴드 마스킹 중의 적어도 하나가 수행되는, 장치.

청구항 12

제11항에 있어서,

상기 미세조정 학습부에 포함된 마스킹부는:

상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치 중 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치를 식별하고;

상기 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치에 대해 마스킹을 수행하는, 장치.

청구항 13

제12항에 있어서,

상기 오디오 신호 입력 데이터와 연관된 샘플링 속도 유형의 수에 기초하여 상기 복수의 패치 중 마스킹되는 패치의 비율 패턴의 수가 결정되는, 장치.

청구항 14

제8항에 있어서,

상기 제1 인코더 및 상기 제2 인코더는, 비전 트랜스포머(vision transformer) 구조를 포함하는, 장치.

발명의 설명

기술 분야

[0001] 본 개시는 오디오 음질 향상을 위한 오디오 신호 처리에 관한 것으로서, 보다 상세하게는 마스크된 오토인코더를 사용하여 오디오 신호의 주파수 대역폭을 확장하는 오디오 초해상도 방법 및 장치에 관한 것이다.

배경 기술

[0002] 아날로그 기반의 오디오 신호를 디지털 기반의 오디오 신호로 저장하는 과정을 수행할 경우, 샘플링 속도에 따라 오디오의 고주파 부분에 해당하는 정보가 손실될 수 있다.

[0003] 고주파수 대역폭을 확장하는 오디오 초해상도는 고주파수 정보가 손실된 주파수 해상도가 낮은 오디오 신호의 고주파수 성분을 복원함으로써 원래의 고해상도 오디오 신호로 복구하는 기술을 의미한다.

[0004] 기존에는 동영상으로부터 저음질의 음성을 추출하여 고음질의 음성으로 변환하고 다시 동영상에 합치는 기술이 고안되었다. 다만, 해당 기술을 구현하기 위해서는 오디오 초해상도 모델을 학습하기 위한 저음질 및 고음질 데이터쌍이 필요하다는 문제점이 존재한다. 이에 따라, 음질이 다른 한 쌍의 오디오 데이터를 수집하기 위한 비용이 추가적으로 발생할 수 있다.

[0005] 또한, 기존에는 보간법 등을 이용한 기존 주파수 대역폭 확장 기술들은 손실된 고주파 성분을 정확하게 복원하는데 한계가 있으며, 아티팩트(artifact) 또는 왜곡이 발생할 수 있다는 문제점이 존재한다.

발명의 내용

해결하려는 과제

- [0006] 본 개시는 상술된 문제점을 해결하기 위해 안출된 것으로서, 본 개시의 목적은 마스크된 오디오인코더를 사용하여 오디오 신호의 주파수 대역폭을 확장하는 오디오 초해상도 방법 및 장치를 제공함에 있다.
- [0007] 또한, 본 개시의 목적은 오디오 신호의 주파수 대역폭을 효과적으로 확장하여 주파수 해상도가 낮은 오디오로부터 주파수 해상도가 높은 고품질의 오디오 신호를 출력하는 방법 및 장치를 제공함에 있다.
- [0008] 본 개시의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 개시의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 개시의 목적 및 장점들은 청구범위에 나타난 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

- [0009] 본 개시의 일 실시예에 따른, 장치에 의해 수행되는, 고주파수 영역의 오디오 신호를 복원하는 방법은, 제1 오디오 신호 학습 데이터에 대해 무작위 마스킹(masking)을 수행한 제1 마스킹 데이터를 제1 인코더(encoder)에 입력하여 제1 인코딩 데이터를 획득하고, 상기 제1 인코딩 데이터를 디코딩한 데이터 및 타겟 데이터 간의 차이에 기초하여 상기 제1 인코더를 학습시키는 사전 학습 단계; 제2 오디오 신호 학습 데이터에 대해 마스킹을 수행한 제2 마스킹 데이터를 제2 인코더에 입력하여 제2 인코딩 데이터를 획득하고, 상기 제2 인코딩 데이터를 오디오 신호 생성부에 입력하여 출력한 오디오 신호 데이터 및 상기 타겟 데이터에 기초하여 상기 제2 인코더 및 상기 오디오 신호 생성부를 학습시키는 미세 조정 학습 단계; 및 제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 상기 학습된 제2 인코더에 입력하여 제3 인코딩 데이터를 획득하고, 상기 제3 인코딩 데이터를 통해 상기 제1 주파수 해상도보다 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득하는 추론 단계를 포함하고, 상기 사전 학습 단계를 통해 학습된 상기 제1 인코더의 가중치는 상기 제2 인코더로 전이될 수 있다.
- [0010] 그리고, 상기 제1 오디오 신호 학습 데이터, 상기 제2 오디오 신호 학습 데이터, 및 상기 오디오 신호 입력 데이터 각각은, 복수의 패치로 분할된 멜-스펙트로그램 형태를 가질 수 있다.
- [0011] 그리고, 제1 오디오 신호 학습 데이터에 대해 무작위 마스킹을 수행하는 단계는, 상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치로부터 무작위로 적어도 하나의 패치를 선택하는 단계; 및 상기 적어도 하나의 패치에 대해 마스킹을 수행하는 단계를 포함할 수 있다.
- [0012] 그리고, 상기 제2 오디오 신호 학습 데이터에 대해 무작위 마스킹 또는 특정 밴드 마스킹 중의 적어도 하나가 수행될 수 있다.
- [0013] 그리고, 상기 제2 오디오 신호 학습 데이터에 대해 특정 밴드 마스킹을 수행하는 단계는, 상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치 중 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치를 식별하는 단계; 상기 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치에 대해 마스킹을 수행하는 단계를 포함할 수 있다.
- [0014] 그리고, 상기 오디오 신호 입력 데이터와 연관된 샘플링 속도 유형의 수에 기초하여 상기 복수의 패치 중 마스킹되는 패치의 비율 패턴의 수가 결정될 수 있다.
- [0015] 그리고, 상기 제1 인코더 및 상기 제2 인코더는, 비전 트랜스포머(vision transformer) 구조를 포함할 수 있다.
- [0016] 본 개시의 일 실시예로, 고주파수 영역의 오디오 신호를 복원하는 장치는, 제1 오디오 신호 학습 데이터에 대해 무작위 마스킹(masking)을 수행한 제1 마스킹 데이터를 제1 인코더(encoder)에 입력하여 제1 인코딩 데이터를 획득하고, 상기 제1 인코딩 데이터를 디코딩한 데이터 및 타겟 데이터 간의 차이에 기초하여 상기 제1 인코더를 학습시키는 사전 학습부; 제2 오디오 신호 학습 데이터에 대해 마스킹을 수행한 제2 마스킹 데이터를 제2 인코더에 입력하여 제2 인코딩 데이터를 획득하고, 상기 제2 인코딩 데이터를 오디오 신호 생성부에 입력하여 출력한 오디오 신호 데이터 및 상기 타겟 데이터에 기초하여 상기 제2 인코더 및 상기 오디오 신호 생성부를 학습시키는 미세 조정 학습부; 및 제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 상기 학습된 제2 인코더에 입력하여 제3 인코딩 데이터를 획득하고, 상기 제3 인코딩 데이터를 통해 상기 제1 주파수 해상도보다 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득하는 추론부를 포함하고, 상기 사전 학습 단계를 통해 학습된 상기 제1 인코더의 가중치는 상기 제2 인코더로 전이될 수 있다.

- [0017] 그리고, 상기 사전 학습부에 포함된 마스크부는: 상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치로부터 무작위로 적어도 하나의 패치를 선택하고; 및 상기 적어도 하나의 패치에 대해 마스크를 수행할 수 있다.
- [0018] 그리고, 상기 미세 조정 학습부에 포함된 마스크부는: 상기 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치 중 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치를 식별하고; 상기 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치에 대해 마스크를 수행할 수 있다.
- [0019] 상기 과제의 해결 수단은, 본 개시의 특징을 모두 열거한 것은 아니다. 본 개시의 다양한 특징과 그에 따른 장점과 효과는 아래의 구체적인 실시예를 참조하여 보다 상세하게 이해될 수 있을 것이다.

발명의 효과

- [0020] 본 개시의 다양한 실시예에 의해, 마스크된 오토인코더를 사용하여 오디오 신호의 주파수 대역폭을 확장하는 오디오 초해상도 방법 및 장치가 제공될 수 있다.
- [0021] 또한, 본 개시의 다양한 실시예에 의해, 오디오 신호의 주파수 대역폭을 효과적으로 확장하여 주파수 해상도가 낮은 오디오로부터 주파수 해상도가 높은 고품질의 오디오 신호를 출력하는 방법 및 장치가 제공될 수 있다.
- [0022] 또한, 본 개시의 다양한 실시예에 의해, 사전에 학습된 모델을 활용하여 일반화 능력을 향상시키고 윗밴드 마스크 방법을 사용하여 효율적으로 모델을 학습시킴으로써, 고품질의 고주파수 대역폭을 효율적으로 확장할 수 있다.
- [0023] 또한, 본 개시의 다양한 실시예에 의해, 무작위 마스크 비율로 마스크를 수행함으로써, 입력 오디오 신호의 샘플링 속도와 상관없이 대역폭 확장이 가능해질 수 있다.
- [0024] 본 개시의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

- [0025] 도 1은 본 개시의 일 실시예에 따른, 고주파수 대역폭을 확장하는 오디오 초해상도 방법을 수행하는 장치의 구성을 설명하기 위한 블록도이다.
- 도 2는 본 개시의 일 실시예에 따른, 사전 학습부, 미세 조정 학습부의 구성을 설명하기 위한 도면이다.
- 도 3은 본 개시의 일 실시예에 따른, 전처리부의 구성 및 동작을 설명하기 위한 도면이다.
- 도 4는 본 개시의 일 실시예에 따른, 마스크부에 포함된 무작위 마스크부, 마스크 비율 무작위 추출부, 및 특정 밴드 마스크부의 동작을 설명하기 위한 도면이다.
- 도 5는 본 개시의 일 실시예에 따른, 추론부의 동작을 설명하기 위한 순서도이다.
- 도 6은 본 개시의 일 실시예에 따른, 고주파수 대역폭을 확장하는 오디오 초해상도 방법을 구현하는 방식을 설명하기 위한 도면이다.
- 도 7은 본 개시의 일 실시예에 따른, 고주파수 영역의 오디오 신호를 복원하는 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 본 개시의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시의 개시가 완전하도록 하고, 본 개시가 속하는 기술 분야의 통상의 기술자에게 본 개시의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 개시는 청구항의 범주에 의해 정의될 뿐이다.
- [0027] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 개시를 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다.

- [0028] 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 개시의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.
- [0029] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 개시가 속하는 기술 분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.
- [0030] 공간적으로 상대적인 용어인 "아래(below)", "아래(beneath)", "하부(lower)", "위(above)", "상부(upper)" 등은 도면에 도시되어 있는 바와 같이 하나의 구성요소와 다른 구성요소와의 상관관계를 용이하게 기술하기 위해 사용될 수 있다. 공간적으로 상대적인 용어는 도면에 도시되어 있는 방향에 더하여 사용시 또는 동작 시 구성요소들의 서로 다른 방향을 포함하는 용어로 이해되어야 한다.
- [0031] 예를 들어, 도면에 도시되어 있는 구성요소를 뒤집을 경우, 다른 구성요소의 "아래(below)"또는 "아래(beneath)"로 기술된 구성요소는 다른 구성요소의 "위(above)"에 놓여질 수 있다. 따라서, 예시적인 용어인 "아래"는 아래와 위의 방향을 모두 포함할 수 있다. 구성요소는 다른 방향으로도 배향될 수 있으며, 이에 따라 공간적으로 상대적인 용어들은 배향에 따라 해석될 수 있다.
- [0032] 이하에서는 도면을 참조하여 마스크된 오토인코더에 기초하여 고주파수 대역폭을 확장하는 오디오 초해상도 방법에 대해 구체적으로 설명하도록 한다.
- [0034] 도 1은 본 개시의 일 실시예에 따른, 고주파수 대역폭을 확장하는 오디오 초해상도 방법을 수행하는 장치의 구성을 설명하기 위한 블록도이다.
- [0035] 도 1에서 설명하는 장치는 스마트 폰, 태블릿 PC, 웨어러블 장치, 노트북 등을 포함할 수 있으나 이에 제한되는 것은 아니다.
- [0036] 도 1에 도시된 바와 같이, 장치(100)는 메모리(110), 통신 모듈(120) 및 프로세서(130)를 포함할 수 있다. 다만, 이에 국한되는 것은 아니며, 장치(100)는 필요한 동작에 따라 당업자 관점에서 자명한 범위 내에서 소프트웨어 및 하드웨어 구성이 수정/추가/생략될 수 있다.
- [0037] 메모리(110)는 본 장치(100)의 다양한 기능을 지원하는 데이터와, 제어부의 동작을 위한 프로그램을 저장할 수 있고, 입/출력되는 데이터들을 저장할 있고, 본 장치에서 구동되는 다수의 응용 프로그램(application program 또는 애플리케이션(application)), 본 장치의 동작을 위한 데이터들, 명령어들을 저장할 수 있다. 이러한 응용 프로그램 중 적어도 일부는, 무선 통신을 통해 외부 장치로부터 다운로드 될 수 있다.
- [0038] 이러한, 메모리(110)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), SSD 타입(Solid State Disk type), SDD 타입(Silicon Disk Drive type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(random access memory; RAM), SRAM(static random access memory), 롬(read-only memory; ROM), EEPROM(electrically erasable programmable read-only memory), PROM(programmable read-only memory), 자기 메모리, 자기 디스크 및 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 또한, 메모리(110)는 본 장치와는 분리되어 있으나, 유선 또는 무선으로 연결된 데이터베이스가 될 수도 있다.
- [0039] 통신 모듈(120)은 외부 장치와 통신을 가능하게 하는 하나 이상의 구성 요소를 포함할 수 있다. 예를 들어, 통신 모듈(120)은 무선통신 모듈, 유선통신 모듈 또는 위치정보 모듈 중 적어도 하나를 포함할 수 있다.
- [0040] 여기서, 무선 통신 모듈은 와이파이(Wifi) 모듈, 와이브로(Wireless broadband) 모듈 외에도, GSM(global System for Mobile Communication), CDMA(Code Division Multiple Access), WCDMA(Wideband Code Division Multiple Access), UMTS(universal mobile telecommunications system), TDMA(Time Division Multiple Access), LTE(Long Term Evolution), 4G, 5G, 6G 등 다양한 무선 통신 방식을 지원하는 무선 통신 모듈을 포함할 수 있다.

- [0041] 프로세서(130)는 본 장치(100) 내의 구성요소들의 동작을 제어하기 위한 알고리즘 또는 알고리즘을 재현한 프로그램에 대한 데이터를 저장하는 메모리, 및 메모리에 저장된 데이터를 이용하여 전술한 동작을 수행하는 적어도 하나의 프로세서(미도시)로 구현될 수 있다. 이때, 메모리와 프로세서는 각각 별개의 칩으로 구현될 수 있다. 또는, 메모리와 프로세서는 단일 칩으로 구현될 수도 있다.
- [0042] 즉, 프로세서(130)는 본 장치(100) 내의 구성 요소들을 제어함으로써 장치의 전반적인 동작 및 기능을 제어할 수 있다. 프로세서는 이하의 도면에서 설명되는 본 개시에 따른 다양한 실시 예들을 본 장치 상에서 구현하기 위하여, 위에서 살펴본 구성요소들을 중 어느 하나 또는 복수를 조합하여 제어할 수 있다.
- [0043] 본 개시의 일 예로, 도 1에 도시된 바와 같이, 프로세서(130)는 사전 학습부(130-1), 미세 조정 학습부(130-2) 및 추론부(130-3)를 포함할 수 있다. 즉, 프로세서(130)는 사전 학습부(130-1), 미세 조정 학습부(130-2) 및 추론부(130-3)를 제어할 수 있다.
- [0044] 다만, 이는 일 실시예에 불과하며, 사전 학습부(130-1), 미세 조정 학습부(130-2) 및 추론부(130-3) 각각은 별개의 장치 또는/및 시스템에 포함될 수 있다. 이 때, 사전 학습부(130-1), 미세 조정 학습부(130-2) 및 추론부(130-3) 각각은 별개의 장치 또는/및 시스템 내에 포함된 통신 모듈을 통해 데이터를 교환할 수 있다.
- [0045] 사전 학습부(130-1)는 일반화 능력이 뛰어난 부호화부(즉, 인코더)를 생성/구현하기 위하여 대용량의 오디오 데이터를 이용하여 인공지능 모델을 학습시킬 수 있다.
- [0046] 사전 학습부(130-1)는 전처리부, 마스킹부, 부호화부, 및 복호화부(즉, 디코더)를 포함할 수 있다. 사전 학습부(130-1)의 세부적인 구성 및 동작은 도 2 내지 도 5를 참조하여 구체적으로 설명하도록 한다.
- [0047] 미세 조정 학습부(130-2)는 추론부(130-3)에서 사용할 인공지능 모델을 학습시킬 수 있다. 미세 조정 학습부(130-2)는 전처리부, 마스킹부, 부호화부, 및 오디오 신호 생성부를 포함할 수 있다. 사전 학습부(130-1)에서 학습된 가중치는 미세 조정 학습부(130-2)의 부호화부로 전이될 수 있다. 이에 따라, 인공지능 모델은 대용량 데이터 및 미세 조정 학습을 통해 일반화 능력 및 고도화 능력이 개선될 수 있다.
- [0048] 미세 조정 학습부(130-2)의 세부적인 구성 및 동작은 도 2 내지 도 5를 참조하여 구체적으로 설명하도록 한다.
- [0049] 추론부(130-3)는 인공지능 모델을 이용하여 입력된 오디오 데이터에 대해 초해상도를 수행할 수 있다. 즉, 추론부(130-3)는 오디오 신호를 사전 학습 및 미세 조정 학습이 수행된 인공지능 모델에 입력하여 해당 오디오 신호에 대해 주파수 대역폭 확장 동작을 수행할 수 있다.
- [0050] 여기서, 추론부(130-3)는 전처리부, 부호화부 및 오디오 신호 생성부를 포함할 수 있다. 미세 조정 학습부(130-2)의 세부적인 구성 및 동작은 도 2 및 도 5를 참조하여 구체적으로 설명하도록 한다.
- [0052] 도 2, 도 3, 및 도 4, 도 5 및 도 6은 본 개시의 일 실시예에 따른, 사전 학습부, 미세 조정 학습부, 및 추론부의 동작을 설명하기 위한 도면이다.
- [0053] 구체적으로, 도 2는 사전 학습부, 미세 조정 학습부의 구성을 설명하기 위한 도면이다. 도 3은 전처리부의 구성 및 동작을 설명하기 위한 도면이다. 도 4는 마스킹부에 포함된 무작위 마스킹부, 마스킹 비율 무작위 추출부, 및 특정 밴드 마스킹부의 동작을 설명하기 위한 도면이다. 도 5는 추론부의 동작을 설명하기 위한 순서도이다.
- [0054] 도 2에 도시된 바와 같이, 사전 학습부(130-1) 및 미세 조정 학습부(130-2) 각각은 전처리부를 포함할 수 있다.
- [0055] 전처리부는 오디오 신호에 대해 단시간 푸리에 변환(short-time Fourier transform) 및 멜 필터 뱅크를 적용하여 중간 표현인 멜-스펙트로그램으로 변환할 수 있다.
- [0056] 도 3에 도시된 바와 같이, 전처리부는 단시간 푸리에 변환부(310), 멜 필터뱅크 적용부(320) 및 패치 분할부(330)를 포함할 수 있다.
- [0057] 구체적으로, 단시간 푸리에 변환부(310)는 입력된 오디오 신호에 대해 단시간 푸리에 변환을 수행함으로써 오디오 신호에 대응되는 스펙트로그램을 생성할 수 있다. 멜 필터뱅크 적용부(320)는 오디오 신호에 대응되는 스펙트로그램에 대해 멜 필터뱅크를 적용하여 멜-스펙트로그램을 생성할 수 있다. 패치 분할부(330)는 멜-스펙트로그램을 서로 겹치지 않는 일정한 크기의 직사각형 또는 정사각형의 패치로 분할한 후 이를 출력할 수 있다.
- [0058] 도 2에 도시된 바와 같이, 사전 학습부(130-1) 및 미세 조정 학습부(130-2) 각각은 마스킹부를 포함할 수 있다.

마스킹부는 전처리부에 의해 생성된 스펙트로그램에 대해 마스킹을 적용할 수 있다. 마스킹부는 무작위 마스킹부, 마스킹 비율 무작위 추출부, 및 특정 밴드(예로, 윗 밴드 또는/및 고주파수 영역) 마스킹부를 포함할 수 있다.

- [0059] 도 4에 도시된 바와 같이, 전처리부에 의해 출력된 중간 표현(즉, 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터)이 추론부에 입력되었는지 여부에 따라 중간 표현의 처리 동작이 결정될 수 있다(S410).
- [0060] 일 예로, 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터가 추론부에 입력된 경우, 추론부는 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터를 부호화부를 통해 인코딩할 수 있다. 그리고, 추론부는 오디오 신호 생성부를 통해 인코딩된 데이터를 인공지능 모델에 입력함으로써 주파수 대역폭이 확장된 오디오 신호를 획득할 수 있다.
- [0061] 또 다른 예로, 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터가 추론부에 입력되지 않은 경우, 해당 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터가 미세 조정 학습부에 입력된 지 여부에 따라 마스킹 방식이 달라질 수 있다(S420).
- [0062] 일 예로, 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터가 미세 조정 학습부에 입력되지 않은 경우(즉, 중간 표현이 사전 학습부에 입력된 경우), 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터는 사전 학습부의 마스킹부에 포함된 무작위 마스킹부에 입력될 수 있다. 구체적으로, 사전 학습부의 마스킹부에 포함된 무작위 마스킹부는 복수의 패치 중 일부를 무작위로 선출한 후 이에 대해 마스킹을 수행할 수 있다.
- [0063] 또 다른 예로, 패치 분할부에 의해 분할된 멜-스펙트로그램 데이터가 미세 조정 학습부에 입력된 경우, 미세 조정 학습부는 무작위 비율을 사용할지 여부를 결정할 수 있다(S430).
- [0064] 구체적으로, 무작위 비율을 사용하지 않을 경우, 미세 조정 학습부에 포함된 특정 밴드 마스킹부는 고주파에 해당하는 특정 밴드의 패치만을 마스킹할 수 있다. 무작위 비율을 사용할 경우, 미세 조정 학습부에 포함된 특정 밴드 마스킹부는 마스킹 비율을 무작위로 설정한 후 설정된 마스킹 비율로 패치들에 대한 마스킹 동작을 수행할 수 있다.
- [0065] 여기서, 다양한 샘플링 속도의 오디오 신호를 입력받는 것으로 결정된 경우, 미세 조정 학습부는 마스킹 비율을 무작위로 설정하는 것으로 결정할 수 있다. 또 다른 예로, 특정 샘플링 속도의 오디오 신호를 입력받는 것으로 결정된 경우, 미세 조정 학습부는 무작위 비율을 특정 비율로 고정할 수 있다.
- [0066] 도 2에 도시된 바와 같이, 사전 학습부(130-1) 및 미세 조정 학습부(130-2) 각각은 부호화부(즉, 인코더)를 포함할 수 있다. 사전 학습부(130-1)의 부호화부에서 학습시킨 가중치는 미세 조정 학습부(130-2)의 부호화부로 전이시켜 학습될 수 있다.
- [0067] 사전 학습부(130-1) 및 미세 조정 학습부(130-2)에 포함된 부호화부는 마스킹된 스펙트로그램을 부호화할 수 있다. 추론부(130-3)의 부호화부는 전처리부를 통해 획득된 스펙트로그램을 부호화할 수 있다.
- [0068] 사전 학습부(130-1)의 복호화부는 부호화된 스펙트로그램을 복원할 수 있다. 미세 조정 학습부(130-2) 및 추론부(130-3)의 오디오 신호 생성부는 부호화된 스펙트로그램을 이용하여 실제 오디오 신호를 생성할 수 있다.
- [0069] 즉, 추론부(130-3)의 오디오 신호 생성부는 사전 학습부(130-1) 및 미세 조정 학습부(130-2)에 의해 학습된 인공지능 모델을 통해 주파수 대역폭이 확장된 실제 오디오 신호(즉, 주파수 해상도가 개선된 오디오 신호)를 획득할 수 있다.
- [0070] 도 5는 본 개시의 일 실시예에 따른, 추론부의 동작을 설명하기 위한 순서도이다.
- [0071] 추론부는 입력된 오디오 신호에 대해 전처리 동작을 수행할 수 있다(S510). 추론부는 입력된 오디오 신호에 대해 단시간 푸리에 변환, 멜 필터 뱅크 적용 및 패치 분할 동작을 수행할 수 있다. 즉, 추론부의 전처리 동작은 도 3을 참조하여 설명한 전처리 동작과 동일할 수 있다.
- [0072] 추론부는 전처리 수행한 데이터에 대해 부호화 동작을 수행할 수 있다(S520). 추론부는 전처리를 수행한 데이터를 사전 학습부 및 미세 조정 학습부에 의해 학습된 부호화부에 입력하여 부호화 동작을 수행할 수 있다.
- [0073] 추론부는 부호화된 데이터에 기초하여 오디오 신호를 생성할 수 있다(S530). 추론부는 부호화된 데이터를 오디오 신호 생성부에 입력하여 오디오 신호를 출력할 수 있다.
- [0074] 즉, 추론부는 사전 학습부 및 미세 조정 학습부에 의해 학습된 부호화부 및 오디오 신호 생성부로 구성된 인공

지능 모델을 통해 높은 주파수 해상도를 가지는 오디오 신호를 출력할 수 있다.

[0075]

[0076]

도 6은 본 개시의 일 실시예에 따른, 고주파수 대역폭을 확장하는 오디오 초해상도 방법을 구현하는 방식을 설명하기 위한 도면이다.

[0077]

도 6에 도시된 바와 같이, 주파수 해상도가 개선된 오디오 신호를 생성하기 위한 인공지능 모델을 학습시키기 위하여, 사전 학습(610) 및 미세 조정 학습(620)이 수행될 수 있다. 사전 학습(610)과 관련된 동작은 상술된 사전학습부에 의해 제어될 수 있으며, 미세 조정 학습(620)과 관련된 동작은 상술된 미세 조정 학습부에 의해 제어될 수 있다.

[0078]

사전 학습(610) 단계의 경우, 사전 학습부는 입력 데이터를 인코더에 입력하여 부호화 동작을 수행할 수 있다. 여기서, 입력 데이터는 오디오 입력 신호에 대해 전처리를 수행한 후 무작위로 마스킹을 수행한 데이터 데이터(즉, 복수의 패치로 분할된 멜-스펙트로그램에 대해 무작위로 마스킹을 수행한 데이터)를 의미할 수 있다.

[0079]

사전 학습부는 전처리를 수행한 데이터를 디코더에 입력하여 멜-스펙트로그램을 복원할 수 있다. 사전 학습부는 멜-스펙트로그램과 타겟 스펙트로그램 간의 차이(즉, MSE(mean square error))에 기초하여 인코더를 학습시킬 수 있다.

[0080]

미세 조정 학습(620) 단계의 경우, 미세 조정 학습부는 입력 데이터를 인코더에 입력하여 부호화 동작을 수행할 수 있다.

[0081]

여기서, 입력 데이터는 오디오 입력 신호에 대해 전처리를 수행한 후 고주파수에 해당하는 패치에 대해서만 마스킹한 데이터를 의미할 수 있다. 그리고, 사전 학습부에 의해 학습된 인코더의 가중치는 미세 조정 학습(620) 단계에서 사용되는 인코더에 전이될 수 있다.

[0082]

미세 조정 학습부에 의해 복호화된 데이터는 오디오 신호 생성부에 입력하여 주파수 해상도가 개선된 오디오 신호를 출력할 수 있다. 미세 조정 학습부는 출력된 오디오 신호에 대해 단시간 푸리에 변환 및 멜 필터를 적용할 수 있다. 미세 조정 학습부는 단시간 푸리에 변환 및 멜 필터가 적용된 데이터와 타겟 스펙트로그램 간의 차이에 기초하여 인코더를 학습시킬 수 있다. 미세 조정 학습부는 출력된 오디오 신호를 판별자(discriminator)에 입력하여 해당 오디오 신호가 실제(real) 개선된 오디오 신호인지 여부를 판별할 수 있다.

[0084]

도 7은 본 개시의 일 실시예에 따른, 고주파수 영역의 오디오 신호를 복원하는 방법을 설명하기 위한 순서도이다.

[0085]

장치는 주파수 해상도 향상을 위한 모델의 일반화 능력을 향상시키기 위하여 사전 학습 단계를 수행할 수 있다(S710).

[0086]

구체적으로, 장치는 제1 오디오 신호 학습 데이터에 대해 무작위 마스킹(masking)을 수행할 수 있다. 여기서, 제1 오디오 신호 학습 데이터는 전처리가 수행된 데이터를 의미할 수 있다. 즉, 도 3을 참조하여 설명한 바와 같이, 제1 오디오 신호 학습 데이터는 복수의 패치로 분할된 멜-스펙트로그램 형태를 가질 수 있다.

[0087]

장치는 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치로부터 무작위로 적어도 하나의 패치를 선택할 수 있다. 장치는 적어도 하나의 패치에 대해 마스킹을 수행할 수 있다.

[0088]

장치는 제1 마스킹 데이터를 제1 인코더(encoder)(즉, 도 1의 사전 학습부의 부호화부)에 입력하여 제1 인코딩 데이터를 획득할 수 있다. 장치는 제1 인코딩 데이터를 디코딩한 데이터 및 타겟 데이터 간의 차이에 기초하여 제1 인코더를 학습시킬 수 있다.

[0089]

여기서, 타겟 데이터는 제1 오디오 신호 학습 데이터에 라벨링된 데이터를 의미할 수 있다. 즉, 타겟 데이터는 장치가 제1 오디오 신호 학습 데이터를 이용하여 획득하고자 하는 주파수 해상도가 개선된 오디오 신호 학습 데이터를 의미할 수 있다.

[0090]

장치는 대역폭 확장 오디오의 초해상도를 수행하도록 모델을 미세 조정 학습 단계를 수행할 수 있다(S720).

[0091]

구체적으로, 장치는 제2 오디오 신호 학습 데이터에 대해 마스킹을 수행하여 제2 마스킹 데이터를 획득할 수 있다. 여기서, 제2 오디오 신호 학습 데이터는 제1 오디오 신호 학습 데이터와 동일할 수 있으나 이에 제한되는

것은 아니다.

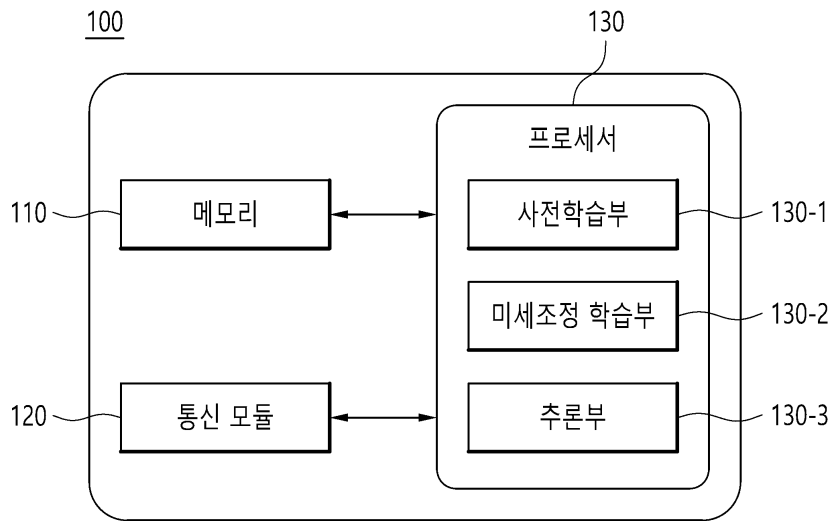
- [0092] 즉, 제2 오디오 신호 학습 데이터는 전처리가 수행된 데이터를 의미할 수 있다. 즉, 도 3을 참조하여 설명한 바와 같이, 제2 오디오 신호 학습 데이터는 복수의 패치로 분할된 멜-스펙트로그램 형태를 가질 수 있다.
- [0093] 장치는 제2 오디오 신호 학습 데이터에 대해 무작위 마스킹 또는 특정 밴드 마스킹 중의 적어도 하나를 수행할 수 있다.
- [0094] 일 예로, 제2 오디오 신호 학습 데이터에 대해 특정 밴드 마스킹을 수행하는 경우를 가정한다.
- [0095] 장치는 제1 오디오 신호 학습 데이터를 구성하는 복수의 패치 중 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치를 식별할 수 있다. 장치는 임계 주파수 이상의 대역에 대응되는 적어도 하나의 패치에 대해 마스킹을 수행할 수 있다.
- [0096] 한편, 추후 추론 단계에서 입력할 오디오 신호 입력 데이터와 연관된 샘플링 속도 유형의 수에 기초하여 복수의 패치 중 마스킹되는 패치의 비율 패턴의 수가 결정될 수 있다.
- [0097] 즉, 추후 다양한 샘플링 속도의 오디오 신호가 입력되는 경우, 다양한 비율로 마스킹 동작이 수행될 수 있다. 특정 샘플링 속도의 오디오 신호가 집중적으로 입력되는 경우, 특정 비율로 마스킹 동작이 수행될 수 있다.
- [0098] 장치는 제2 인코더에 입력하여 제2 인코딩 데이터를 획득할 수 있다. 이 때, 사전 학습 단계를 통해 학습된 제1 인코더의 가중치는 제2 인코더로 전이된 상태일 수 있다. 그리고, 제1 인코더 및 제2 인코더는 비전 트랜스포머(vision transformer) 구조를 포함할 수 있다.
- [0099] 장치는 제2 인코딩 데이터를 오디오 신호 생성부에 입력하여 출력한 오디오 신호 데이터 및 타겟 데이터에 기초하여 제2 인코더 및 오디오 신호 생성부를 학습시킬 수 있다.
- [0100] 장치는 제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 통해 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득하는 추론 단계를 수행할 수 있다(S730).
- [0101] 구체적으로, 장치는 제1 주파수 해상도를 가지는 오디오 신호 입력 데이터를 미세 조정 학습 단계를 통해 학습된 제2 인코더에 입력하여 제3 인코딩 데이터를 획득할 수 있다. 그리고, 장치는 제3 인코딩 데이터를 미세 조정 학습 단계를 통해 학습된 오디오 신호 생성부에 입력하여 제1 주파수 해상도보다 높은 제2 주파수 해상도를 가지는 오디오 신호 출력 데이터를 획득할 수 있다.
- [0103] 이상, 본 개시에서 설명되는 실시예와 첨부된 도면은 본 개시에 포함되는 기술적 사상의 일부를 예시적으로 설명하는 것에 불과하다. 따라서, 본 명세서에 개시된 실시예들은 본 개시의 기술적 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이므로, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것이 아님은 자명하다.
- [0104] 본 개시의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시 예는 모두 본 개시의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

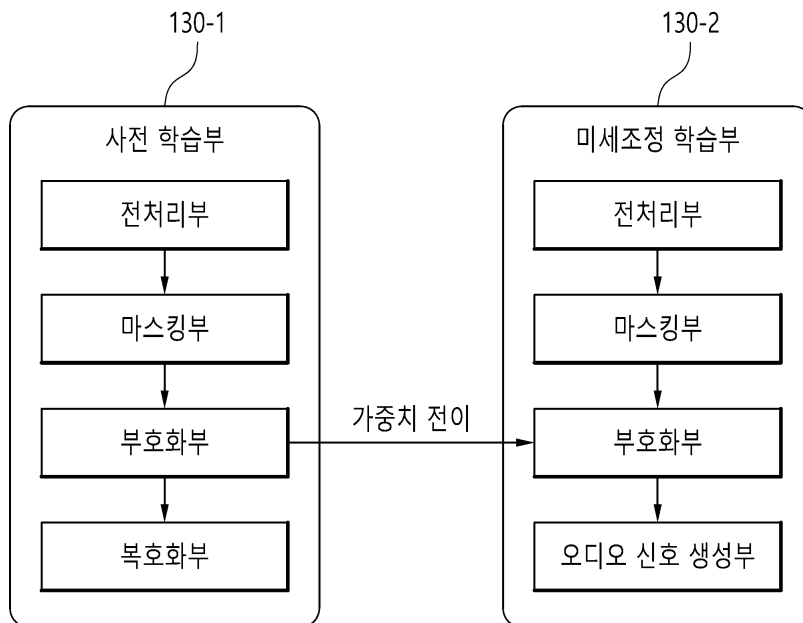
- [0106] 100 : 장치
- 110 : 메모리
- 120 : 통신 모듈
- 130 : 프로세서
- 130-1 : 사전 학습부
- 130-2 : 미세 조정 학습부
- 130-3 : 추론부

도면

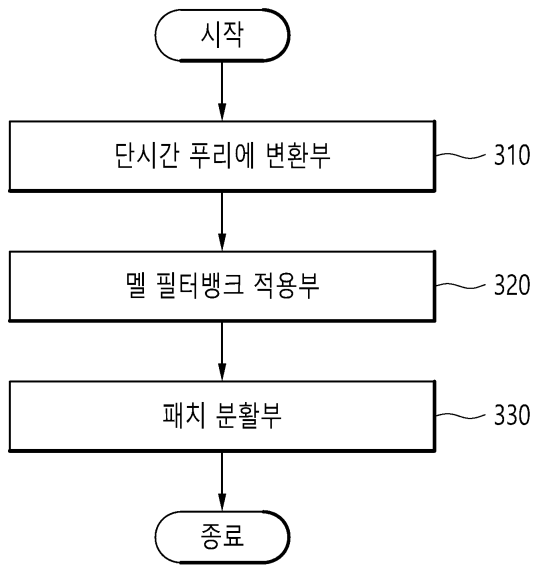
도면1



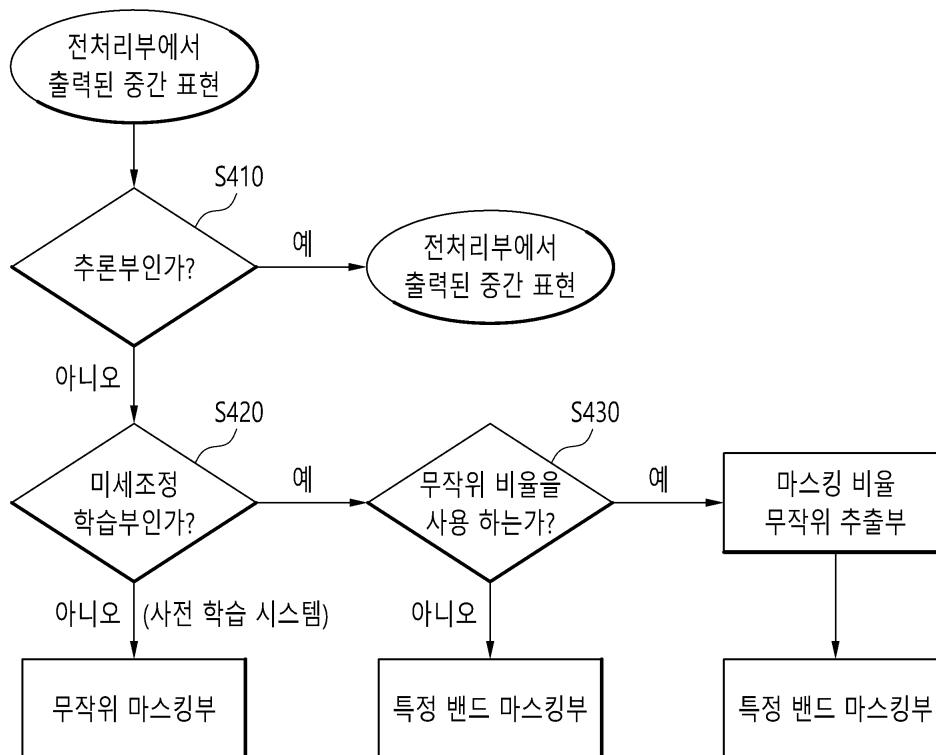
도면2



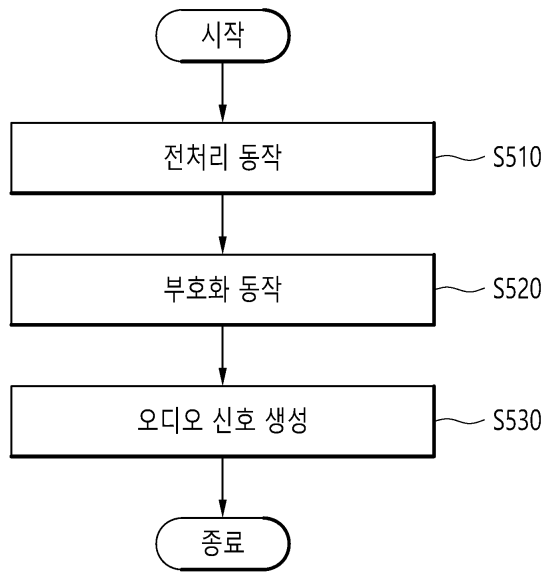
도면3



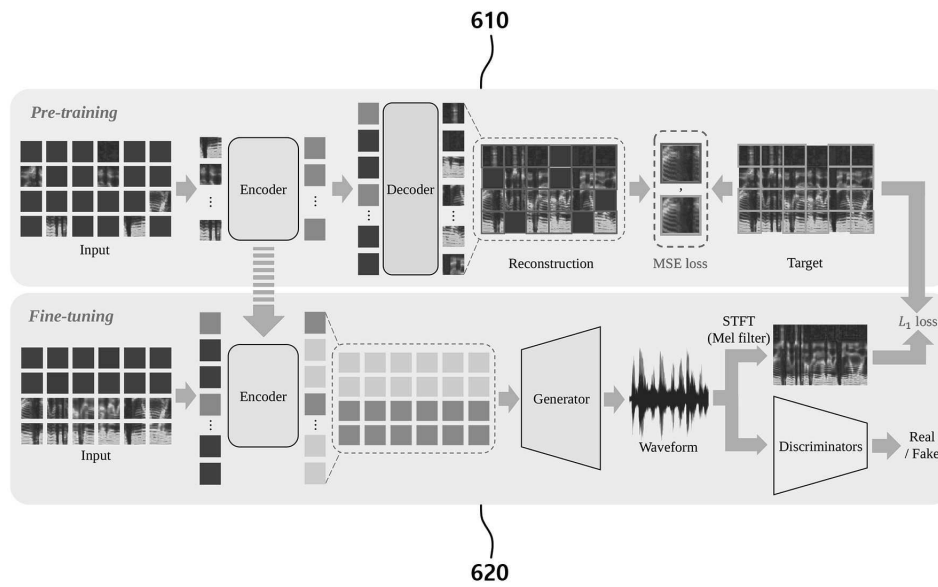
도면4



도면5



도면6



도면7

