



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2022년05월19일

(11) 등록번호 10-2400598

(24) 등록일자 2022년05월17일

(51) 국제특허분류(Int. Cl.)
G10L 21/0208 (2013.01) **G06N 3/08** (2006.01)
G10L 25/30 (2013.01)

(52) CPC특허분류
G10L 21/0208 (2013.01)
G06N 3/08 (2013.01)

(21) 출원번호 10-2020-0084272

(22) 출원일자 2020년07월08일

심사청구일자 2020년07월08일

(65) 공개번호 10-2022-0006382

(43) 공개일자 2022년01월17일

(56) 선행기술조사문헌

임경현 외 2명, '비정상 잡음 제거를 위한 이중
 관별자 적대적 생성모델', 2019년 한국소프트웨
 어종합학술대회 논문집, 2019년 12월 18일.*

Kyung-hyun Lim et al., 'Non-stationary noise
 cancellation using deep autoencoder based on
 adversarial learning', IDEAL 2019, LNCS
 11871, pp. 367-374, November 2019.*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

연세대학교 산학협력단

서울특별시 서대문구 연세로 50 (신촌동, 연세대
 학교)

(72) 발명자

조성배

서울특별시 강남구 선릉로76길 12, 101동 201호(
 대치동, 대치한신휴플러스)

김진영

경기도 하남시 미사강변대로 165, 110동 1303호(
 망월동, 미사강변 푸르지오)

임경현

서울특별시 양천구 오목로 299, B동 3703호(목동,
 트라펠리스웨스턴에비뉴)

(74) 대리인

특허법인우인

전체 청구항 수 : 총 11 항

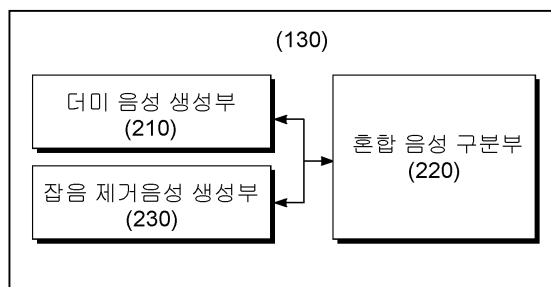
심사관 : 정성윤

(54) 발명의 명칭 기계학습 기반의 잡음 제거 방법 및 그를 위한 장치

(57) 요약

기계학습 기반의 잡음 제거 방법 및 그를 위한 장치를 개시한다.

본 발명의 실시예에 따른 잡음 제거 방법은, 화자의 음성과 잡음으로 구성된 혼합 음성을 입력 받는 음성 입력
 단계; 상기 혼합 음성을 기반으로 실제 원본 음성인 제2 음성(A)과 비교하기 위한 제1 음성(A')을 생성하는 더미
 음성 생성 단계; 상기 제1 음성을 상기 제2 음성(A)을 이용하여 음성을 구분하고, 상기 혼합 음성과 상기 제1 음
 성 및 상기 제2 음성 각각을 기반으로 생성된 잔차 잡음을 이용하여 잡음을 구분하여 학습하는 혼합 음성 구분
 단계; 및 상기 음성 및 상기 잡음을 구분한 학습결과를 기반으로 신규 혼합 음성에 대한 잡음을 제거한 최종 음
 성을 생성하는 잡음 제거음성 생성 단계를 포함할 수 있다.

대표도

(52) CPC특허분류
G10L 25/30 (2013.01)

공지예외적용 : 있음

명세서

청구범위

청구항 1

하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 컴퓨팅 디바이스에 의해 수행되는 잡음 제거 방법에 있어서, 상기 컴퓨팅 디바이스는,

화자의 음성과 잡음으로 구성된 혼합 음성을 입력 받는 음성 입력 단계;

상기 혼합 음성을 기반으로 실제 원본 음성인 제2 음성(A)과 비교하기 위한 제1 음성(A')을 생성하는 더미 음성 생성 단계;

상기 제1 음성을 상기 제2 음성(A)을 이용하여 음성을 구분하고, 상기 혼합 음성과 상기 제1 음성 및 상기 제2 음성 각각을 기반으로 생성된 잔차 잡음을 이용하여 잡음을 구분하여 학습하는 혼합 음성 구분 단계; 및

상기 음성 및 상기 잡음을 구분한 학습결과를 기반으로 신규 혼합 음성에 대한 잡음을 제거한 최종 음성을 생성하는 잡음 제거음성 생성 단계를 수행하되,

상기 혼합 음성 구분 단계는, 상기 제1 음성 및 상기 제2 음성이 동일한 음성인지 여부를 구분하는 음성 구분 단계; 및 상기 혼합 음성과 상기 제1 음성을 기반으로 생성된 제1 잔차 잡음과 상기 혼합 음성과 상기 제2 음성을 기반으로 생성된 제2 잔차 잡음이 동일한 잡음인지 여부를 구분하는 잡음 구분 단계를 포함하는 것을 특징으로 하는 잡음 제거 방법.

청구항 2

제1항에 있어서,

상기 더미 음성 생성 단계는,

상기 실제 원본 음성의 웨이브폼 세그먼트와 여러 잡음들이 합성된 상기 혼합 음성의 웨이브폼 세그먼트가 매핑된 데이터셋을 사용하여 상기 제1 음성을 생성하는 것을 특징으로 하는 잡음 제거 방법.

청구항 3

제1항에 있어서,

상기 더미 음성 생성 단계는,

상기 혼합 음성에서 잡음을 제거하기 위하여 상기 혼합 음성의 압축을 수행하는 혼합 음성 압축 단계; 및

압축된 혼합 음성을 재구성하여 상기 제1 음성(A')을 생성하는 데이터 재구성 단계

를 포함하는 것을 특징으로 하는 잡음 제거 방법.

청구항 4

삭제

청구항 5

제1항에 있어서,

상기 음성 구분 단계는,

상기 제1 음성과 상기 제2 음성을 입력 받고, 상기 제1 음성이 상기 제2 음성과 동일한 음성인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(flag) 값을 출력하는 것을 특징으로 하는 잡음 제거 방법.

청구항 6

제1항에 있어서,

상기 잡음 구분 단계는,

상기 혼합 음성과 상기 제1 음성을 기반으로 생성된 제1 잔차 잡음과 상기 혼합 음성과 상기 제2 음성을 기반으로 생성된 제2 잔차 잡음을 입력 받고, 상기 제1 잔차 잡음이 상기 제2 잔차 잡음과 동일한 음성인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(flag) 값을 출력하는 것을 특징으로 하는 잡음 제거 방법.

청구항 7

제6항에 있어서,

상기 잡음 구분 단계는,

상기 혼합 음성에서 상기 제1 음성을 제거한 상기 제1 잔차 잡음과 상기 혼합 음성에서 상기 제2 음성을 제거한 상기 제2 잔차 잡음을 비교하여 잡음을 구분하는 것을 특징으로 하는 잡음 제거 방법.

청구항 8

제1항에 있어서,

상기 음성 구분 단계는,

상기 더미 음성 생성 단계와 연동하여 상기 제1 음성 및 상기 제2 음성을 구분하기 위하여 생성적 적대 신경망(GAN: Generative Adversarial Network) 학습을 수행하며,

상기 잡음 구분 단계는, 상기 더미 음성 생성 단계와 연동하여 상기 제1 잔차 잡음 및 상기 제2 잔차 잡음을 구분하기 위하여 생성적 적대 신경망(GAN) 학습을 수행하는 것을 특징으로 하는 잡음 제거 방법.

청구항 9

제1항에 있어서,

상기 잡음 제거음성 생성 단계는,

상기 음성 구분 단계와 연동하여 상기 제1 음성과 상기 제2 음성을 구분한 제1 학습 결과를 획득하고, 상기 잡음 구분 단계와 연동하여 상기 제1 잔차 잡음과 상기 제2 잔차 잡음을 구분한 제2 학습 결과를 획득하며, 상기 제1 학습 결과와 상기 제2 학습 결과를 기반으로 잡음이 제거된 최종 음성을 생성하는 것을 특징으로 하는 잡음 제거 방법.

청구항 10

혼합 음성에서 잡음을 제거하는 장치로서,

하나 이상의 프로세서; 및

상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로그램들은 하나 이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서들에서,

화자의 음성과 잡음으로 구성된 혼합 음성을 입력 받는 음성 입력 단계;

상기 혼합 음성을 기반으로 실제 원본 음성인 제2 음성(A)과 비교하기 위한 제1 음성(A')을 생성하는 더미 음성 생성 단계;

상기 제1 음성을 상기 제2 음성(A)을 이용하여 음성을 구분하고, 상기 혼합 음성과 상기 제1 음성 및 상기 제2 음성 각각을 기반으로 생성된 잔차 잡음을 이용하여 잡음을 구분하여 학습하는 혼합 음성 구분 단계; 및

상기 음성 및 상기 잡음을 구분한 학습결과를 기반으로 신규 혼합 음성에 대한 잡음을 제거한 최종 음성을 생성하는 잡음 제거음성 생성 단계를 포함하는 동작들을 수행하게 하되,

상기 혼합 음성 구분 단계는, 상기 제1 음성 및 상기 제2 음성이 동일한 음성인지 여부를 구분하는 음성 구분 단계; 및 상기 혼합 음성과 상기 제1 음성을 기반으로 생성된 제1 잔차 잡음과 상기 혼합 음성과 상기 제2 음성을 기반으로 생성된 제2 잔차 잡음이 동일한 잡음인지 여부를 구분하는 잡음 구분 단계를 포함하는 것을 특징으로 하는 잡음 제거 장치.

청구항 11

삭제

청구항 12

제10항에 있어서,

상기 음성 구분 단계는,

상기 제1 음성과 상기 제2 음성을 입력 받고, 상기 제1 음성이 상기 제2 음성과 동일한 음성인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(flag) 값을 출력하는 것을 특징으로 하는 잡음 제거 장치.

청구항 13

제10항에 있어서,

상기 잡음 구분 단계는,

상기 혼합 음성과 상기 제1 음성을 기반으로 생성된 제1 잔차 잡음과 상기 혼합 음성과 상기 제2 음성을 기반으로 생성된 제2 잔차 잡음을 입력 받고, 상기 제1 잔차 잡음이 상기 제2 잔차 잡음과 동일한 음성인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(flag) 값을 출력하는 것을 특징으로 하는 잡음 제거 장치.

발명의 설명

기술 분야

[0001] 본 발명은 기계학습을 사용하여 음성신호에서 잡음을 제거하는 방법 및 그를 위한 장치에 관한 것이다.

배경 기술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 잡음 제거 시스템은 잡음이 없는 음성과 잡음이 포함된 단일 채널의 음성 신호를 입력으로 받아 잡음을 제거한 후 음성만 남아있는 신호를 생성하는 방법을 학습한 모델을 적용한 시스템을 의미한다. 여기서, 잡음 제거는 신호 처리 연구의 한 분야로써, 전화, 음성 통신 및 토론 상황 등 여러 잡음이 섞여 화자 음성 내용의 이해가 어려울 수 있는 상황과 같이 잡음이 많은 복잡한 환경에서 잡음을 제거하여 깨끗한 음성을 추출할 때 유용하게 활용될 수 있으며, 많은 분야에서 응용되어 사용할 수 있기 때문에 연구가 활발히 진행되고 있다.

[0004] 하지만 신호 중에서도 음성과 같이 연속적인 시계열 데이터의 형태를 하고 다양한 피치(pitch), 주파수를 가진 신호들의 분포를 학습하고 수많은 잡음 패턴을 학습하여 잡음만을 제거하고 단채널을 통해 들어오는 하나의 입력만으로 원하는 타겟 음성만을 추출하는 것은 어렵다.

[0005] 특히, 지도 학습 (supervised learning)을 이용하는 시계열 데이터 처리를 위한 모델을 기반으로 한 기존 잡음 제거 신경망 방법들의 한계가 드러나고 있기 때문에 이를 해결하기 위한 방법이 요구된다. 또한 적대 학습 방법은 생성 모델이 넓은 데이터의 분포를 학습 과정에서 쉽게 망가질 수 있다는 문제를 가지고 있기 때문에 효과적으로 잡음을 제거하는 모델을 만들기 위한 안정적인 학습 방법이 요구된다.

[0006] 기존의 잡음 제거 연구에서는 잡음이 포함되지 않은 음성 소스(x)와 잡음 소스(n)들이 섞인 잡음 음성($y = x + n$)을 지도학습 방법을 사용하여 분포를 매핑(mapping)하는 시도를 해왔다. 이 학습 과정은 잡음이 포함된 음성들의 근본이 되는 모든 음성 즉, 잡음이 포함된 음성과 잡음이 포함되지 않는 음성을 1:1로 구축한 데이터로부터 큰 분포를 학습해야 한다는 어려움이 존재한다.

[0007] 또한, 기존의 많은 연구들에서는 시퀀스 데이터를 처리하기 위하여 RNN(Recurrent Neural Network)과 같은 순환 모델들을 주로 사용했고, 이런 모델들의 고질적인 문제점인 그래디언트 베니싱 문제, 즉 긴 시퀀스 데이터에는 대응하기 어렵다는 문제점이 있으며, 잡음을 제거하는 과정에서도 문제점이 발생한다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 두 개의 생성자(generator)와 두 개의 구분자(discriminator)를 이용해 하나의 생성자를 적대 학습 방법을 통해 훈련시킴으로써 잡음이 포함된 음성에서 잡음을 제거하는 성능을 향상시키고 잡음이 포함된 음성과 생성자로부터 생성된 음성과의 잔차를 이용하여 잡음을 제거하기 위한 적대 학습의 효과를 향상시키고 그 학습 과정을 안정화 시키는 기계학습 기반의 잡음 제거 방법 및 그를 위한 장치를 제공하는 데 주된 목적이 있다.

과제의 해결 수단

[0009] 본 발명의 일 측면에 의하면, 상기 목적을 달성하기 위한 잡음 제거 방법은, 화자의 음성과 잡음으로 구성된 혼합 음성을 입력 받는 음성 입력 단계; 상기 혼합 음성을 기반으로 실제 원본 음성인 제2 음성(A)과 비교하기 위한 제1 음성(A')을 생성하는 더미 음성 생성 단계; 상기 제1 음성을 상기 제2 음성(A)을 이용하여 음성을 구분하고, 상기 혼합 음성과 상기 제1 음성 및 상기 제2 음성 각각을 기반으로 생성된 잔차 잡음을 이용하여 잡음을 구분하여 학습하는 혼합 음성 구분 단계; 및 상기 음성 및 상기 잡음을 구분한 학습결과를 기반으로 신규 혼합 음성에 대한 잡음을 제거한 최종 음성을 생성하는 잡음 제거음성 생성 단계를 포함할 수 있다.

[0010] 또한, 본 발명의 다른 측면에 의하면, 상기 목적을 달성하기 위한 잡음 제거 장치는, 하나 이상의 프로세서; 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로그램들은 하나 이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서들에서, 화자의 음성과 잡음으로 구성된 혼합 음성을 입력 받는 음성 입력 단계; 상기 혼합 음성을 기반으로 실제 원본 음성인 제2 음성(A)과 비교하기 위한 제1 음성(A')을 생성하는 더미 음성 생성 단계; 상기 제1 음성을 상기 제2 음성(A)을 이용하여 음성을 구분하고, 상기 혼합 음성과 상기 제1 음성 및 상기 제2 음성 각각을 기반으로 생성된 잔차 잡음을 이용하여 잡음을 구분하여 학습하는 혼합 음성 구분 단계; 및 상기 음성 및 상기 잡음을 구분한 학습결과를 기반으로 신규 혼합 음성에 대한 잡음을 제거한 최종 음성을 생성하는 잡음 제거음성 생성 단계를 수행할 수 있다.

발명의 효과

[0011] 이상에서 설명한 바와 같이, 본 발명은 잔차를 이용한 적대 학습 기반의 잡음 제거 시스템 및 음성 생성 학습 방법으로 일반적인 학습 과정에 비해 잡음 제거 성능을 증대시킬 수 있는 효과가 있다.

[0012] 또한, 본 발명은 잡음 구분부와 적대 학습을 통해 더미 음성 생성부가 원본 음성을 제외한 잡음들을 제거하는 성능을 향상시킬 수 있는 효과가 있다.

[0013] 또한, 본 발명은 기존의 지도 학습 기반 방법들과 달리 잡음이 포함되지 않은 음성과 잡음이 포함된 음성이 서로 쌍을 이루어 정렬된 데이터가 필수적으로 존재하지 않아도 학습이 가능하며, 다른 딥러닝 기반의 방법들과 다르게 별도의 전처리 작업없이 음성 신호를 자연 그대로 사용해 정보의 손실을 최소화할 수 있는 효과가 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 실시예에 따른 잡음 제거 장치를 개략적으로 나타낸 블록 구성도이다.

도 2는 본 발명의 실시예에 따른 프로세서의 동작 구성을 나타낸 블록 구성도이다.

도 3은 본 발명의 실시예에 따른 더미 음성 생성부 및 잡음 제거음성 생성부의 동작 구성을 나타낸 블록 구성도이다.

도 4a 및 도 4b는 본 발명의 실시예에 따른 데이터 압축 및 데이터 재구성을 위한 신경망 구조를 나타낸 도면이다.

도 5는 본 발명의 실시예에 따른 구분부의 동작 구성을 나타낸 블록 구성도이다.

도 6 및 도 7은 본 발명의 실시예에 따른 음성 구분부 및 잡음 구분부의 신경망 구조를 나타낸 도면이다.

도 8은 본 발명의 실시예에 따른 잡음 제거 방법을 설명하기 위한 순서도이다.

도 9는 본 발명의 실시예에 따른 생성적 적대 신경망 기반의 잡음 제거 동작을 설명하기 위한 예시도이다.

도 10a 및 도 10b는 본 발명의 실시예에 따른 잡음 제거 장치의 학습 과정을 설명하기 위한 도면이다.

도 11은 본 발명의 실시예에 따른 잡음 제거 장치의 적용 과정을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0015] 이하, 본 발명의 바람직한 실시예를 첨부된 도면들을 참조하여 상세히 설명한다. 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다. 또한, 이하에서 본 발명의 바람직한 실시예를 설명할 것이나, 본 발명의 기술적 사상은 이에 한정하거나 제한되지 않고 당업자에 의해 변형되어 다양하게 실시될 수 있음은 물론이다. 이하에서는 도면들을 참조하여 본 발명에서 제안하는 기계학습 기반의 잡음 제거 방법 및 그를 위한 장치에 대해 자세하게 설명하기로 한다.
- [0016] 도 1은 본 발명의 실시예에 따른 잡음 제거 장치를 개략적으로 나타낸 블록 구성도이다.
- [0017] 본 실시예에 따른 잡음 제거 장치(100)는 입력부(110), 출력부(120), 프로세서(130), 메모리(140) 및 데이터 베이스(150)를 포함한다. 도 1의 잡음 제거 장치(100)는 일 실시예에 따른 것으로서, 도 1에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 잡음 제거 장치(100)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 잡음 제거 장치(100)는 컴퓨팅 디바이스로 구현될 수 있고, 잡음 제거 장치(100)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 장치로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.
- [0018] 잡음 제거 장치(100)는 잡음이 포함된 혼합 음성을 입력으로 받아 원본 음성을 제외한 잡음을 모두 제거한 음성으로 출력하는 모델 및 잔차 기반 적대 학습을 통한 잡음 제거 모델을 구축하여 혼합 음성의 잡음을 제거하는 동작을 수행한다.
- [0019] 입력부(110)는 잡음 제거 장치(100)의 잡음 제거 동작을 수행하기 위한 신호 또는 데이터를 입력하거나 획득하는 수단을 의미한다. 입력부(110)는 프로세서(130)와 연동하여 다양한 형태의 신호 또는 데이터를 입력하거나, 외부 장치와 연동하여 직접 데이터를 획득하여 프로세서(130)로 전달할 수도 있다. 여기서, 입력부(110)는 혼합 음성, 특정 화자의 음성, 잡음 등을 입력하기 위한 마이크로 구현될 수 있으나 반드시 이에 한정되는 것은 아니다.
- [0020] 출력부(120)는 프로세서(130)와 연동하여 혼합 음성의 잡음 제거 결과, 학습 결과, 최종 음성 결과 등 다양한 정보를 표시할 수 있다. 출력부(120)는 잡음 제거 장치(100)에 구비된 디스플레이(미도시)를 통해 다양한 정보를 표시하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다.
- [0021] 프로세서(130)는 메모리(140)에 포함된 적어도 하나의 명령어 또는 프로그램을 실행시키는 기능을 수행한다.
- [0022] 본 실시예에 따른 프로세서(130)는 입력부(110) 또는 데이터 베이스(150)로부터 획득한 혼합 음성을 기반으로 기계학습을 수행하고, 기계학습 결과를 기반으로 혼합 음성에서 화자의 음성, 잡음 등을 분리하고, 이를 통해 신규 혼합 음성에서 잡음을 제거하는 동작을 수행한다.
- [0023] 프로세서(130)는 혼합 음성을 입력 받고, 혼합 음성을 기반으로 특정 화자의 실제 원본 음성과 비교하여 학습하기 위한 제1 음성을 생성하고, 생성된 제1 음성을 특정 화자의 실제 원본 음성인 제2 음성(A)과 구분하고, 혼합 음성과 제1 음성 및 상기 제2 음성을 기반으로 생성된 잔차 잡음을 구분하는 학습 동작을 수행하여 혼합 음성에 대한 잡음 제거가 수행되도록 한다. 본 실시예에 따른 프로세서(130)의 자세한 동작은 도 2에서 설명하도록 한다.
- [0024] 메모리(140)는 프로세서(130)에 의해 실행 가능한 적어도 하나의 명령어 또는 프로그램을 포함한다. 메모리(140)는 음성을 생성하는 동작, 음성을 구분하는 동작, 잡음을 구분하는 동작, 잡음을 제거하는 동작 등을 위한 명령어 또는 프로그램을 포함할 수 있다. 또한, 메모리(140)는 학습 결과를 적용하는 동작, 화자를 분리하는 동작 등을 위한 명령어 또는 프로그램을 포함할 수 있다.
- [0025] 데이터 베이스(150)는 데이터베이스 관리 프로그램(DBMS)을 이용하여 컴퓨터 시스템의 저장공간(하드디스크 또는 메모리)에 구현된 일반적인 데이터구조를 의미하는 것으로, 데이터의 검색(추출), 삭제, 편집, 추가 등을 자유롭게 행할 수 있는 데이터 저장형태를 뜻하는 것으로, 오라클(Oracle), 인포믹스(Infomix), 사이베이스(Sybase), DB2와 같은 관계형 데이터베이스 관리 시스템(RDBMS)이나, 겔스톤(Gemston), 오리온(Orion), O2 등과 같은 객체 지향 데이터베이스 관리 시스템(OODBMS) 및 엑셀론(Excelon), 타미노(Tamino), 세카이주(Sekaiju) 등의 XML 전용 데이터베이스(XML Native Database)를 이용하여 본 발명의 일 실시예의 목적에 맞게 구현될 수 있고, 자신의 기능을 달성하기 위하여 적당한 필드(Field) 또는 엘리먼트들을 가지고 있다.

- [0026] 본 실시예에 따른 데이터베이스(400)는 잡음 제거와 관련된 데이터를 저장하고, 잡음 제거와 관련된 데이터를 제공할 수 있다.
- [0027] 데이터베이스(400)에 저장된 데이터는 혼합 음성, 특정 화자의 음성, 잡음, 학습 결과, 최종 음성 등에 대한 데이터일 수 있다. 데이터베이스(140)는 잡음 제거 장치(100) 내에 구현되는 것으로 기재하고 있으나 반드시 이에 한정되는 것은 아니며, 별도의 데이터 저장장치로 구현될 수도 있다.
- [0028] 도 2는 본 발명의 실시예에 따른 프로세서의 동작 구성을 나타낸 블록 구성도이다.
- [0029] 본 실시예에 따른 잡음 제거 장치(100)에 포함된 프로세서(130)는 기계 학습을 기반으로 잡음 제거를 처리하는 동작을 수행한다. 여기서, 기계 학습은 생성적 적대 신경망(GAN: Generative Adversarial Network)을 이용한 학습인 것이 바람직하나 반드시 이에 한정되는 것은 아니다.
- [0030] 잡음 제거 장치(100)에 포함된 프로세서(130)는 잡음이 포함된 혼합 음성을 입력으로 받아 원본 음성을 제외한 잡음을 모두 제거한 음성으로 출력하는 모델(음성 구분부) 및 잔차 기반 적대 학습을 통한 잡음 제거 모델(잡음 구분부) 구축 방법을 기반으로 동작하며, 잡음 제거를 수행해야 하는 모든 기기 및 소프트웨어에 탑재될 수 있다. 예를 들어, 잡음 제거 장치(100)에 포함된 프로세서(130)는 AI 스피커, 스마트폰과 같은 음성 인식 기술이 접목되어 있는 기기 등에 적용되어, 소음이 심한 주변 환경에서도 소유주의 음성만을 구분하는 기능에 응용되어 쓰일 수 있다.
- [0031] 본 실시예에 따른 프로세서(130)는 더미 음성 생성부(210), 혼합 음성 구분부(220) 및 잡음 제거음성 생성부(230)를 포함한다.
- [0032] 더미 음성 생성부(210)는 잡음이 포함된 혼합 음성을 입력 받고, 혼합 음성을 기반으로 화자의 실제 음성인 원본 음성과 비교하여 학습하기 위한 제1 음성을 생성한다. 구체적으로, 더미 음성 생성부(210)는 잡음과 특정 화자의 음성이 결합된 혼합 음성을 입력 받고, 혼합 음성을 기반으로 특정 화자에 대한 제1 음성(A')을 생성하여 출력한다.
- [0033] 더미 음성 생성부(210)는 하나의 실제 원본 음성의 웨이브폼 세그먼트와 여러 잡음들이 합성된 혼합 음성의 웨이브폼 세그먼트가 매핑된 데이터셋을 사용하여 생성된 제1 음성을 출력한다. 더미 음성 생성부(210)의 입력에 들어가는 혼합 음성은 일반화 성능 향상을 위해 실제 음성에 가우시안 분포를 따르는 랜덤 값이 첨가된 음성일 수 있다.
- [0034] 본 실시예에 따른 더미 음성 생성부(210)는 생성적 적대 신경망(GAN)을 기반으로 학습하기 위한 생성자(Generator)로 구현될 수 있으나 반드시 이에 한정되는 것은 아니다.
- [0035] 혼합 음성 구분부(220)는 더미 음성 생성부(210)에서 생성된 제1 음성을 특정 화자의 실제 음성(원본 음성)인 제2 음성(A)과 구분하고, 혼합 음성과 제1 음성 및 제2 음성 각각을 기반으로 생성된 잔차 잡음을 구분하여 잡음 제거가 수행되도록 한다.
- [0036] 혼합 음성 구분부(220)는 더미 음성 생성부(210)에서 생성된 제1 음성을 특정 화자의 실제 음성인 제2 음성(A)과 구분하는 음성 구분부(510)와 혼합 음성과 제1 음성 및 제2 음성 각각을 기반으로 생성된 잔차 잡음을 구분하는 잡음 구분부(520)를 포함한다. 혼합 음성 구분부(220)에 포함된 음성 구분부(510) 및 잡음 구분부(520)은 도 5에서 설명하도록 한다.
- [0037] 본 실시예에 따른 혼합 음성 구분부(220)는 생성적 적대 신경망(GAN)을 기반으로 학습하기 위한 서로 다른 두 개의 구분자(Discriminator)를 포함하는 형태로 구현될 수 있으나 반드시 이에 한정되는 것은 아니다.
- [0038] 혼합 음성 구분부(220)는 더미 음성 생성부(210)와 생성적 적대 신경망(GAN) 기반의 학습을 수행한다. 여기서, 혼합 음성 구분부(220)에 포함된 음성 구분부(510) 및 잡음 구분부(520) 각각은 더미 음성 생성부(210)와 별도로 생성적 적대 신경망(GAN) 기반의 학습을 수행하여 서로 다른 학습 결과에 대한 모델을 생성할 수 있다.
- [0039] 본 발명의 생성적 적대 신경망(GAN)에서, 더미 음성 생성부(210)는 혼합 음성 구분부(220)에서 실제 원본 음성(제2 음성)과 구분할 수 없는 음성(제1 음성)을 생성하는 것을 목표로 한다. 한편, 혼합 음성 구분부(220)의 음성 구분부(510)는 더미 음성 생성부(210)가 생성한 음성(제1 음성)을 실제 원본 음성(제2 음성)과 구분할 수 있도록 하여, 더미 음성 생성부(210)와 혼합 음성 구분부(220)가 적대적으로 학습하는 방식을 말한다.
- [0040] 본 발명의 적대 학습은 음성 생성부(230)에서 실제 원본 음성과 유사한 잡음이 제거된 음성을 생성하는 것을 목표로 한다. 혼합 음성 구분부(220)에서는 더미 음성 생성부(210)가 생성한 음성(제1 음성)과 원본 음성(제2 음

성)을 구별할 수 있도록 하여, 더미 음성 생성부(210)는 음성 구분부(510)와 잡음 구분부(510)를 속이기 위해 어떤 음성을 생성해야 원본 음성(제2 음성)과 유사할 수 있는지에 대해 학습을 한다. 이후 학습된 음성 구분부(510) 및 잡음 구분부(520)에 일종의 교사의 역할을 부여하여 음성 생성부(230)가 안정적으로 실제 음성과 유사한 분포를 학습하는 방식을 말한다.

[0041] 더미 음성 생성부(210)와 혼합 음성 구분부(220)와의 적대 학습은 [수학식 1]과 같이 나타낼 수 있다.

수학식 1

[0042]
$$\min_{G_d} \max_D V(D, G_d) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log (1 - D(G_d(z)))]$$

[0043] 여기서, Pdata(x)는 잡음이 포함되지 않은 원본 음성 데이터의 분포, x는 Pdata(x)로부터 추출한 샘플, P(z)는 기 정의된 사전 분포, z는 P(z)로부터 추출한 샘플, G_d는 더미 음성 생성부(Generator), D는 음성 구분부(Discriminator)를 의미한다.

[0044] 일반적인 적대 학습 방법은 잡음 제거 문제에서 기존의 연구들에 비해 향상된 성능의 생성자를 훈련시키는 것이 어렵기 때문에 본 발명의 혼합 음성 구분부(220)에서는 잔차를 기반으로 한 잡음 구분부(520)를 추가로 포함한다. 잡음 구분부(520)는 본 발명에서 적대 학습을 통해 음성 생성부의 생성능력을 향상 시키기 위해 추가되어 잡음이 포함되지 않은 원본 음성과 잡음이 포함된 음성의 차이를 이용해 잡음을 학습하는 구분자이다.

[0045] 잡음 제거음성 생성부(230)는 생성적 적대 신경망(GAN)의 학습 결과를 기반으로 잡음 제거를 수행하고, 잡음이 제거된 최종 음성을 생성하여 출력한다.

[0046] 잡음 제거음성 생성부(230)는 제1 음성과 제2 음성을 구분한 제1 학습 결과와 제1 잔차 잡음과 제2 잔차 잡음을 구분한 제2 학습 결과를 기반으로 잡음이 제거된 최종 음성을 생성할 수 있다.

[0047] 잡음 제거음성 생성부(230)는 음성 구분부(510)로부터 제1 음성과 제2 음성을 구분한 제1 학습 결과를 획득하고, 잡음 구분부(520)로부터 제1 잔차 잡음과 제2 잔차 잡음을 구분한 제2 학습 결과를 획득한다.

[0048] 이후, 잡음 제거음성 생성부(230)는 제1 학습 결과와 제2 학습 결과를 기반으로 기 저장된 혼합 음성 또는 신규로 입력된 혼합 음성에 포함된 잡음을 제거한 최종 음성을 생성한다.

[0049] 도 3은 본 발명의 실시예에 따른 더미 음성 생성부 및 잡음 제거음성 생성부의 동작 구성을 나타낸 블록 구성도이다.

[0050] 본 실시예에 따른 더미 음성 생성부(210)는 제1 데이터 압축부(310) 및 제1 데이터 재구성부(312)를 포함한다. 도 3의 더미 음성 생성부(210)는 일 실시예에 따른 것으로서, 도 3에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 더미 음성 생성부(210)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다.

[0051] 더미 음성 생성부(210)는 잡음이 포함된 혼합 음성을 입력 받고, 혼합 음성을 기반으로 화자의 실제 음성인 원본 음성과 비교하여 학습하기 위한 제1 음성을 생성한다. 구체적으로, 더미 음성 생성부(210)는 잡음과 특정 화자의 음성이 결합된 혼합 음성을 입력 받고, 혼합 음성을 기반으로 특정 화자에 대한 제1 음성(A')을 생성하여 출력한다. 여기서, 혼합 음성은 특정 화자의 원본 음성과 잡음이 포함된 음성일 수 있으며, 잡음은 다른 화자의 음성, 주변 소리, 기계음 등일 수 있다.

[0052] 더미 음성 생성부(210)는 하나의 실제 원본 음성의 웨이브폼 세그먼트와 여러 잡음들이 합성된 혼합 음성의 웨이브폼 세그먼트가 매핑된 데이터셋을 사용하여 학습을 통해 생성된 제1 음성을 출력한다.

[0053] 제1 데이터 압축부(310)는 혼합 음성에서 잡음을 제거하기 위하여 혼합 음성의 압축을 수행하는 동작을 수행한다.

[0054] 제1 데이터 압축부(310)는 혼합 음성을 압축함으로써 음성의 특징을 추출하여 잠재 공간에 전사할 수 있다. 제1 데이터 압축부(310)는 혼합 음성에서 잡음을 제거(분리)하도록 하기 위하여 혼합 음성으로 압축된 표현형 데이터로 표현할 수 있다.

[0055] 제1 데이터 재구성부(312)는 압축된 혼합 음성을 재구성하여 제1 음성(A')을 생성한다.

[0056] 제1 데이터 재구성부(312)는 압축된 혼합 음성이 실제 원본 음성(제2 음성)과 유사하도록 재구성한다. 즉, 제1

데이터 재구성부(312)는 제1 데이터 압축부(310)로부터 생성된 표현형 데이터를 이용하여 실제 원본 음성을 재구성한 제1 음성(A')을 생성하여 출력한다.

- [0057] 더미 음성 생성부(210)는 기본적으로 오토인코더(AE: AutoEncoder)의 구조로 구현될 수 있다. 예를 들어, 더미 음성 생성부(210)에서 제1 데이터 압축부(310)는 오토인코더(AE)의 인코더(Encoder)와 대응되는 동작을 수행하고, 제1 데이터 재구성부(312)는 오토인코더(AE)의 디코더(Decoder)에 대응되는 동작을 수행할 수 있다.
- [0058] 본 실시예에 따른 잡음 제거음성 생성부(230)는 제2 데이터 압축부(320) 및 제2 데이터 재구성부(322)를 포함한다. 제2 데이터 압축부(320) 및 제2 데이터 재구성부(322)의 구성은 더미 음성 생성부(210)와 동일할 수 있다.
- [0059] 잡음 제거음성 생성부(230)는 음성 구분부(510)로부터 제1 음성과 제2 음성을 구분한 제1 학습 결과를 획득하고, 잡음 구분부(520)로부터 제1 잔차 잡음과 제2 잔차 잡음을 구분한 제2 학습 결과를 획득한다. 잡음 제거음성 생성부(230)는 제1 학습 결과와 제2 학습 결과를 기반으로 기 저장된 혼합 음성 또는 신규로 입력된 혼합 음성에 포함된 잡음을 제거한 최종 음성을 생성한다.
- [0060] 제2 데이터 압축부(320)는 기 저장된 혼합 음성 또는 신규로 입력된 혼합 음성에서 잡음을 제거하기 위하여 혼합 음성의 압축을 수행하는 동작을 수행한다.
- [0061] 제2 데이터 압축부(320)는 혼합 음성을 압축함으로써 음성의 특징을 추출하여 잠재 공간에 전사할 수 있다. 제1 데이터 압축부(310)는 혼합 음성에서 잡음을 제거(분리)하도록 하기 위하여 혼합 음성으로 압축된 표현형 데이터로 표현할 수 있다.
- [0062] 제2 데이터 재구성부(322)는 압축된 혼합 음성을 재구성하여 잡음이 제거된 최종 음성을 생성한다.
- [0063] 제2 데이터 재구성부(322)는 잡음이 포함된 음성을 처리하는데 있어 다양한 잡음에 대해 반응할 수 있게 만들기 위해 원본 음성과 함께 정규 분포를 따르는 랜덤 샘플을 추가할 수 있다.
- [0064] 제2 데이터 재구성부(322)는 압축된 혼합 음성이 실제 원본 음성과 유사하도록 재구성한다. 즉, 제2 데이터 재구성부(322)는 제1 학습 결과와 제2 학습 결과를 기반으로 제2 데이터 압축부(320)로부터 생성된 표현형 데이터를 이용하여 잡음이 제거된 형태를 재구성한 최종 음성을 생성하여 출력한다.
- [0065] 잡음 제거음성 생성부(230)는 기본적으로 오토인코더(AE: AutoEncoder)의 구조로 구현될 수 있다. 예를 들어, 잡음 제거음성 생성부(230)에서 제2 데이터 압축부(320)는 오토인코더(AE)의 인코더(Encoder)와 대응되는 동작을 수행하고, 제2 데이터 재구성부(322)는 오토인코더(AE)의 디코더(Decoder)에 대응되는 동작을 수행할 수 있다.
- [0066] 도 4a 및 도 4b는 본 발명의 실시예에 따른 데이터 압축 및 데이터 재구성을 위한 신경망 구조를 나타낸 도면이다.
- [0067] 도 4a를 참조하면, 제1 데이터 압축부(310)는 인코더에 대응되는 동작을 수행할 수 있다. 예를 들어, 제1 데이터 압축부(310)는 16 KHz 샘플링(sampling)된 모노(mono) 음성파일(raw signal/1초 단위, 16384x1길이)인 잡음이 포함된 혼합 음성(410)을 입력 받는다.
- [0068] 제1 데이터 압축부(310)는 입력 신호(410)를 복수의 서로 다른 크기의 필터(420)로 처리하여 인코딩 벡터(430)를 출력한다. 여기서, 인코딩 벡터(430)는 8x1024 길이로 표현된 벡터일 수 있다.
- [0069] 도 4b를 참조하면, 제1 데이터 재구성부(312)는 디코더에 대응되는 동작을 수행할 수 있다. 예를 들어, 제1 데이터 재구성부(312)는 인코딩 벡터(430)와 가우시안 노이즈(Gaussian noise, 440)를 입력 받는다. 여기서, 인코딩 벡터(430)와 가우시안 노이즈(440)는 동일한 길이의 벡터로 표현될 수 있으며, 8x1024 길이일 수 있다.
- [0070] 제1 데이터 재구성부(312)는 가우시안 노이즈(440)를 추가함으로써 확률적 효과를 부여 학습하지 않은 유사한 잡음에 대한 강건성을 유도할 수 있다.
- [0071] 제1 데이터 재구성부(312)는 인코딩 벡터(430) 및 가우시안 노이즈(440)를 복수의 서로 다른 크기의 필터(450)로 처리하여 제1 음성을 생성한다. 여기서, 생성된 제1 음성은 16384x1 길이로 표현된 벡터일 수 있다.
- [0072] 도 5는 본 발명의 실시예에 따른 구분부의 동작 구성을 나타낸 블록 구성도이다.
- [0073] 본 실시예에 따른 혼합 음성 구분부(220)는 음성 구분부(510) 및 잡음 구분부(520)를 포함한다. 도 4의 혼합 음성 구분부(220)는 일 실시예에 따른 것으로서, 도 4에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시

예에서 혼합 음성 구분부(220)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다.

- [0074] 혼합 음성 구분부(220)는 더미 음성 생성부(210)에서 생성된 제1 음성(A')을 특정 화자의 실제 원본 음성인 제2 음성(A)과 구분하고, 혼합 음성과 제1 음성 및 제2 음성을 기반으로 생성된 잔차 잡음을 구분하여 잡음 제거가 수행되도록 한다.
- [0075] 혼합 음성 구분부(220)는 생성적 적대 신경망(GAN) 학습에 이용되어 더미 음성 생성부(210)에서 생성된 음성과 실제 원본 음성을 구분하고, 혼합 음성에서 더미 음성 생성부(210)가 생성한 음성을 뺀 잔차 잡음과 혼합 음성에서 실제 원본 음성을 뺀 잔차 잡음을 구분하는 동작을 수행한다.
- [0076] 음성 구분부(510)는 제1 음성(생성된 음성) 및 제2 음성(실제 원본 음성)이 동일한 음성인지 여부를 구분하는 동작을 수행한다. 구체적으로, 음성 구분부(510)는 제1 음성과 제2 음성을 입력 받고, 제1 음성이 제2 음성과 동일한 음성인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(Flag) 값을 출력한다.
- [0077] 음성 구분부(510)는 더미 음성 생성부(210)와 연동하여 제1 음성 및 제2 음성을 구분하기 위하여 생성적 적대 신경망(GAN) 학습을 수행할 수 있다.
- [0078] 음성 구분부(510)는 생성적 적대 신경망(GAN) 학습을 통해 더미 음성 생성부(210)에서 재구성되어 생성된 제1 음성과 실제 원본 음성인 제2 음성을 구분하는 성능을 점점 더 향상시킴으로써, 고도화된 음성 구분부(510)를 속이려는 더미 음성 생성부(210)의 제1 음성(생성 음성)의 생성 성능을 향상시킬 수 있다.
- [0079] 잡음 구분부(520)는 혼합 음성과 제1 음성을 기반으로 생성된 제1 잔차 잡음과 혼합 음성과 제2 음성을 기반으로 생성된 제2 잔차 잡음이 동일한 잡음인지 여부를 구분하는 동작을 수행한다.
- [0080] 잡음 구분부(520)는 혼합 음성과 제1 음성을 기반으로 생성된 제1 잔차 잡음과 혼합 음성과 제2 음성을 기반으로 생성된 제2 잔차 잡음을 입력 받고, 제1 잔차 잡음이 제2 잔차 잡음과 동일한 잡음인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(Flag) 값을 출력한다. 여기서, 제1 잔차 잡음은 혼합 음성에서 제1 음성을 제거한 잡음을 의미하고, 제2 잔차 잡음은 혼합 음성에서 제2 음성을 제거한 잡음을 의미한다.
- [0081] 잡음 구분부(520)는 더미 음성 생성부(210)와 연동하여 제1 잔차 잡음 및 제2 잔차 잡음을 구분하기 위하여 생성적 적대 신경망(GAN) 학습을 수행할 수 있다. 잡음 구분부(520)는 잔차를 이용하여 생성적 적대 신경망(GAN) 학습의 효과를 극대화시킬 수 있다.
- [0082] 잡음 구분부(520)는 생성적 적대 신경망(GAN) 학습을 통해 혼합 음성에서 제1 음성을 뺀 제1 잔차 잡음과 혼합 음성에서 제2 음성을 뺀 제2 잔차 잡음을 구분하는 성능을 점점 더 향상시킴으로써, 고도화된 잡음 구분부(520)를 속이려는 더미 음성 생성부(210)의 제1 음성(생성 음성)에 대한 생성 성능을 향상시킬 수 있다.
- [0083] 도 6 및 도 7은 본 발명의 실시예에 따른 음성 구분부 및 잡음 구분부의 신경망 구조를 나타낸 도면이다.
- [0084] 도 6은 음성 구분부(510)의 신경망 구조를 나타낸다. 예를 들어, 음성 구분부(510)는 16 KHz 샘플링(sampling)된 모노(mono) 음성 파일(raw signal/1초 단위, 16384x1 길이)인 실제 원본 음성 또는 더미 음성 생성부(210)에서 생성된 제1 음성을 입력 신호(610)로 획득한다.
- [0085] 음성 구분부(510)는 입력 신호(610)를 복수의 서로 다른 크기의 필터(620)로 처리하여 참 신호 또는 거짓 신호에 대한 플래그(Flag) 값(630)을 출력한다. 여기서, 참 신호는 입력 신호(610)가 실제 원본 음성인 것으로 판단된 경우를 의미하며, True(1) 값으로 표현될 수 있다. 한편, 거짓 신호는 입력 신호(610)가 생성된 음성인 것으로 판단된 경우를 의미하며, False(0) 값으로 표현될 수 있다.
- [0086] 도 7은 잡음 구분부(520)의 신경망 구조를 나타낸다. 예를 들어, 잡음 구분부(520)는 제1 잔차 잡음 또는 제2 잔차 잡음을 입력 신호(710)로 획득할 수 있다. 여기서, 제1 잔차 잡음은 혼합 음성에서 제1 음성을 제거한 생성된 잡음을 의미하고, 제2 잔차 잡음은 혼합 음성에서 제2 음성을 제거한 원본 잡음을 의미한다.
- [0087] 잡음 구분부(520)는 입력 신호(710)를 복수의 서로 다른 크기의 필터(720)로 처리하여 참 신호 또는 거짓 신호에 대한 플래그(Flag) 값(730)을 출력한다. 여기서, 참 신호는 입력 신호(710)가 원본 잡음인 것으로 판단된 경우를 의미하며, True(1) 값으로 표현될 수 있다. 한편, 거짓 신호는 입력 신호(710)가 생성된 잡음인 것으로 판단된 경우를 의미하며, False(0) 값으로 표현될 수 있다.
- [0088] 도 8은 본 발명의 실시예에 따른 잡음 제거 방법을 설명하기 위한 순서도이다.
- [0089] 잡음 제거 장치(100)는 화자의 음성으로 구성된 실제 원본 음성과 잡음이 포함된 혼합음성을 입력 받는다

(S810).

- [0090] 잡음 제거 장치(100)는 혼합음성을 기반으로 실제 원본 음성과 유사한 제1 음성을 생성한다(S820). 잡음 제거 장치(100)는 하나의 실제 원본 음성의 웨이브폼 세그먼트와 여러 잡음들이 합성된 혼합 음성의 웨이브폼 세그먼트가 매핑된 데이터셋을 사용하여 제1 음성을 생성한다.
- [0091] 잡음 제거 장치(100)는 생성적 적대 신경망(GAN) 학습에 이용되어 생성된 제1 음성과 특정 화자의 실제 원본 음성인 제2 음성을 구분한다(S830). 잡음 제거 장치(100)는 제1 음성과 제2 음성을 입력 받고, 제1 음성이 제2 음성과 동일한 음성인지 여부를 구분하여 참 신호 또는 거짓 신호에 대한 플래그(Flag) 값을 출력한다.
- [0092] 잡음 제거 장치(100)는 잔차를 이용하여 혼합 음성에서 제1 음성을 뺀 제1 잔차 잡음과 혼합 음성에서 제2 음성을 뺀 제2 잔차 잡음을 구분한다(S840, S850).
- [0093] 잡음 제거 장치(100)는 생성적 적대 신경망(GAN) 학습을 기반으로 잡음 제거를 수행하고, 잡음이 제거된 최종 음성을 생성하여 출력한다(S860). 잡음 제거 장치(100)는 제1 음성과 제2 음성을 구분한 제1 학습 결과와 제1 잔차 잡음과 제2 잔차 잡음을 구분한 제2 학습 결과를 기반으로 잡음이 제거된 최종 음성을 생성할 수 있다.
- [0094] 도 8에서는 각 단계를 순차적으로 실행하는 것으로 기재하고 있으나, 반드시 이에 한정되는 것은 아니다. 다시 말해, 도 8에 기재된 단계를 변경하여 실행하거나 하나 이상의 단계를 병렬적으로 실행하는 것으로 적용 가능할 것이므로, 도 8은 시계열적인 순서로 한정되는 것은 아니다.
- [0095] 도 8에 기재된 본 실시예에 따른 잡음 제거 방법은 애플리케이션(또는 프로그램)으로 구현되고 단말장치(또는 컴퓨터)로 읽을 수 있는 기록매체에 기록될 수 있다. 본 실시예에 따른 잡음 제거 방법을 구현하기 위한 애플리케이션(또는 프로그램)이 기록되고 단말장치(또는 컴퓨터)가 읽을 수 있는 기록매체는 컴퓨팅 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치 또는 매체를 포함한다.
- [0096] 도 9는 본 발명의 실시예에 따른 생성적 적대 신경망 기반의 잡음 제거 동작을 설명하기 위한 예시도이다.
- [0097] 도 9를 참조하면, 잡음 제거 장치(100)에서 더미 음성 생성부(210)와 혼합 음성 구분부(220) 내의 음성 구분부(510) 및 잡음 구분부(520) 각각은 생성적 적대 신경망(GAN) 학습을 수행한다. 잡음 제거 장치(100)에서 학습된 결과는 잡음 제거음성 생성부(230)로 전달되어, 신규 혼합 음성에 대한 잡음을 제거한 최종 음성이 생성되도록 한다.
- [0098] 더미 음성 생성부(210)는 혼합 음성 구분부(220)에서 실제 원본 음성(제2 음성)과 구분할 수 없는 음성(제1 음성)을 생성하는 것을 목표로 한다. 또한, 혼합 음성 구분부(220)는 더미 음성 생성부(210)에서 생성된 음성(제1 음성)을 실제 원본 음성(제2 음성)과 구분하는 것을 목표로 한다.
- [0099] 생성적 적대 신경망(GAN) 학습이 반복적으로 이루어지면서, 음성 구분부(510)와 잡음 구분부(520)의 구분 성능은 점점 더 향상될 것이고, 더미 음성 생성부(210) 역시 점점 고도화된 음성 구분부(510)와 잡음 구분부(520)를 속이기 위해 재생성 성능이 향상된다.
- [0100] 또한, 잔차를 이용한 잡음 구분부(520)는 생성적 적대 신경망(GAN) 학습의 효과를 극대화 시킨다.
- [0101] 음성 구분부(510)는 더미 음성 생성부(210)에서 생성된 음성(제1 음성)과 실제 원본 음성(제2 음성)을 구분하고, 잡음 구분부(520)는 혼합 음성에서 제1 음성을 제외한 제1 잔차 잡음과 혼합 음성에서 제2 음성을 제외한 제2 잔차 잡음을 이용하여 잡음을 구분하는 역할을 수행한다.
- [0102] 잡음 제거 장치(100)는 화자의 음성을 구분하는 음성 구분부(510)와 잔차를 이용하여 잡음을 구분하는 잡음 구분부(520)를 포함하여 두 번의 생성적 적대 신경망(GAN) 학습이 이루어지게 됨에 따라, 더미 음성 생성부(210)의 제1 음성(생성 음성)의 생성 성능이 강화되고, 혼합 음성에서 화자의 음성을 제외한 잡음들이 더 완벽하게 제거될 수 있다.
- [0104] 도 9를 참고하면, 음성 구분부(510)는 음성에 대한 참 값이 쌍으로 들어오면 참 신호로 판단하고, 거짓 값이 쌍으로 입력되면 거짓 신호로 판단하여 학습하고, 음성 구분부(510)의 목적 함수는 [수학식 2]와 같이 정의될 수 있다.

수학식 2

$$\max_{G_\theta} \min_{D_x} [\{D_x(T(x, \tilde{x})) - 1\}^2 + \{D_x(T(G_\theta(x), \tilde{x}))\}^2] + \|G_\theta(Z, \tilde{x}) - x\|_1$$

또한, 잡음 구분부(520)는 잡음에 대한 참 값이 쌍으로 들어오면 참 신호로 판단하고, 거짓 값이 쌍으로 입력되면 거짓 신호로 판단하여 학습하고, 잡음 구분부(520)의 손실 함수는 [수학식 3]와 같이 정의될 수 있다.

수학식 3

$$\max_{G_\theta} \min_{D_x} [\{D_x(\tilde{x} - x) - 1\}^2 + \{D_x(\tilde{x} - G_\theta(\tilde{x}))\}^2] + \|\{G_\theta(Z, \tilde{x}) - x\} - (\tilde{x} - x)\|_1$$

여기서, x 는 실제 원본 음성을 의미하며 모델이 만들어야 할 대상입니다. \tilde{x} 는 잡음 신호를 의미하고, G_θ 는 더미 음성 생성부(210) 또는 잡음 제거음성 생성부(230) 중 하나를 의미한다. T 는 결합(concatenation)을 의미한다.

잡음 신호는 인코더(잡음 제거음성 생성부(230)의 인코더)의 입력으로 사용되고, 인코더의 출력과 이미 구분된 Z (혼합 음성 구분부(220)의 학습 결과)는 디코더(잡음 제거음성 생성부(230)의 디코더)의 입력으로 사용된다.

도 10a 및 도 10b는 본 발명의 실시예에 따른 잡음 제거 장치의 학습 과정을 설명하기 위한 도면이다.

도 10a의 (a)는 더미 음성 생성부(210)의 학습을 위한 데이터 샘플링을 수행하는 단계(Phase 1)를 나타내고, 도 10a의 (b)는 더미 음성 생성부(210)의 학습을 수행하는 단계(Phase 2)를 나타낸다.

더미 음성 생성부(210)의 역할은 샘플링된 데이터로부터 하이 바운더리(High boundary)와 로우 바운더리(low boundary)를 학습하여 생성부 및 구분부의 학습에 안정성을 부여하는 것이다. 도 10a의 (a)의 과정은 일반적으로 생성적 적대 신경망(GAN) 학습에 불안정한 요소가 있기 때문에 이를 보완하기 위한 과정을 의미한다.

도 10b의 (a)는 구분부에 실제 샘플 학습을 수행하는 단계(Phase 1)를 나타내고, 도 10b의 (b)는 구분부에 가짜 샘플 학습을 수행하는 단계(Phase 2)를 나타내고, 도 10b의 (c)는 음성 생성부에 학습을 수행하는 단계(Phase 3)를 나타낸다.

도 10b의 (a)를 참고하면, 잡음 제거 장치(100)는 음성 구분부(510) 및 잡음 구분부(520) 각각의 구분 결과값이 참 신호가 되도록 학습을 수행한다(Phase 1). 음성 구분부(510)는 원본 음성(제2 음성)을 입력 신호로 학습을 수행하고, 잡음 구분부(520)는 원본 잡음(제2 잔차 잡음)을 입력 신호로 학습을 수행한다.

도 10b의 (b)를 참고하면, 잡음 제거 장치(100)는 음성 구분부(510) 및 잡음 구분부(520) 각각의 구분 결과값이 거짓 신호가 되도록 학습을 수행한다(Phase 2). 음성 구분부(510)는 생성된 음성(제1 음성)을 입력 신호로 학습을 수행하고, 잡음 구분부(520)는 생성된 잡음(제1 잔차 잡음)을 입력 신호로 학습을 수행한다.

도 10b의 (c)를 참고하면, 잡음 제거 장치(100)는 음성 구분부(510) 및 잡음 구분부(520) 각각의 가중치를 고정해두고 음성 생성부만 학습을 수행한다(Phase 3). 즉, 음성 구분부(510) 및 잡음 구분부(520)가 참 신호로 구분할 수 있도록 음성을 생성하기 위하여 음성 생성부를 학습한다.

잡음 제거 장치(100)의 학습 과정에서, 음성 생성부는 잡음 제거음성 생성부(230)일 수 있으며, 잡음 제거음성 생성부(230)의 불안정성을 보완하기 위해 더미 음성 생성부(210)가 학습 가중치를 보조하는 역할을 수행한다. 여기서, 잡음 제거음성 생성부(230)의 불안정성은 구분부(220)를 속이기 쉬운 이상한 파형만을 만들어내는 현상을 의미한다.

도 11은 본 발명의 실시예에 따른 잡음 제거 장치의 적용 과정을 설명하기 위한 도면이다.

잡음 제거 장치(100)는 구분부에 대한 학습이 종료된 이후 잡음 제거 적용시, 잡음 제거음성 생성부(230)만을 사용하여 잡음이 포함된 혼합음성에서 잡음을 제거한다. 여기서, 잡음 제거음성 생성부(230)는 잡음이 포함된 신규 혼합 음성과 가우시안 노이즈를 입력으로 사용하여 동작하며, 잡음이 제거된 최종 음성을 출력한다.

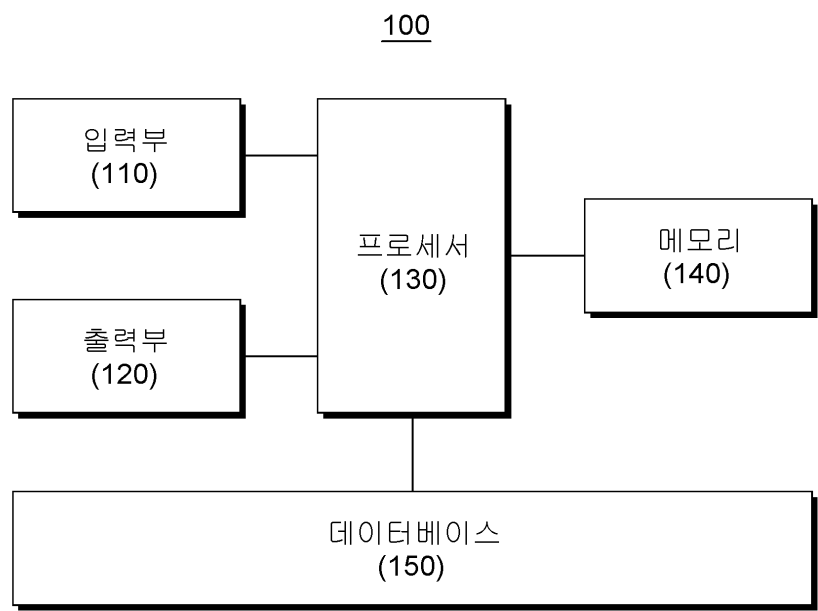
[0120] 이상의 설명은 본 발명의 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명의 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명의 실시예들은 본 발명의 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

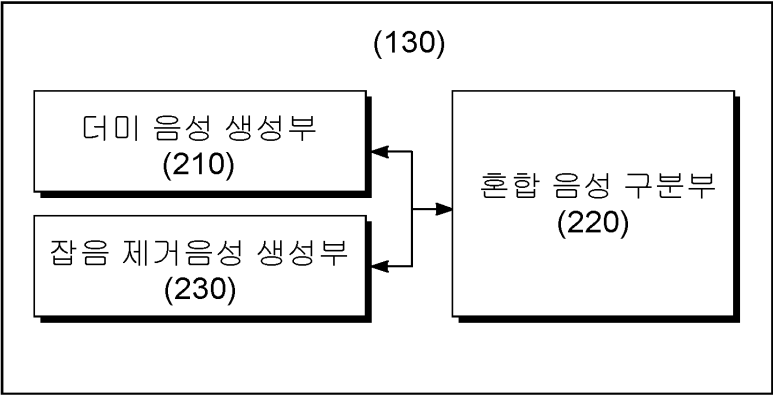
- [0121]
- | | |
|------------------|------------------|
| 100: 잡음 제거 장치 | |
| 110: 입력부 | 120: 출력부 |
| 130: 프로세서 | 140: 메모리 |
| 150: 데이터 베이스 | |
| 210: 더미 음성 생성부 | 220: 혼합 음성 구분부 |
| 230: 잡음 제거음성 생성부 | |
| 310: 제1 데이터 압축부 | 312: 제1 데이터 재구성부 |
| 320: 제2 데이터 압축부 | 322: 제2 데이터 재구성부 |
| 510: 음성 구분부 | 520: 잡음 구분부 |

도면

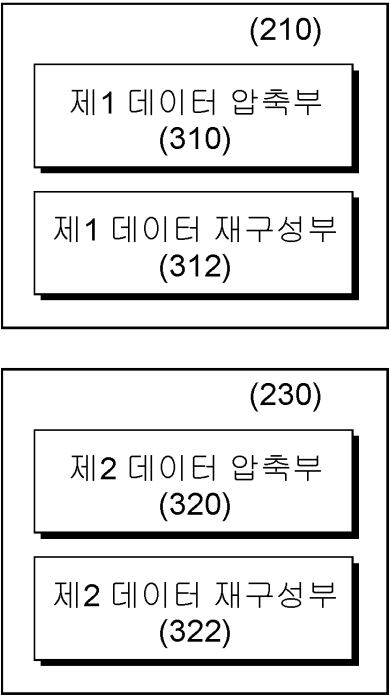
도면1



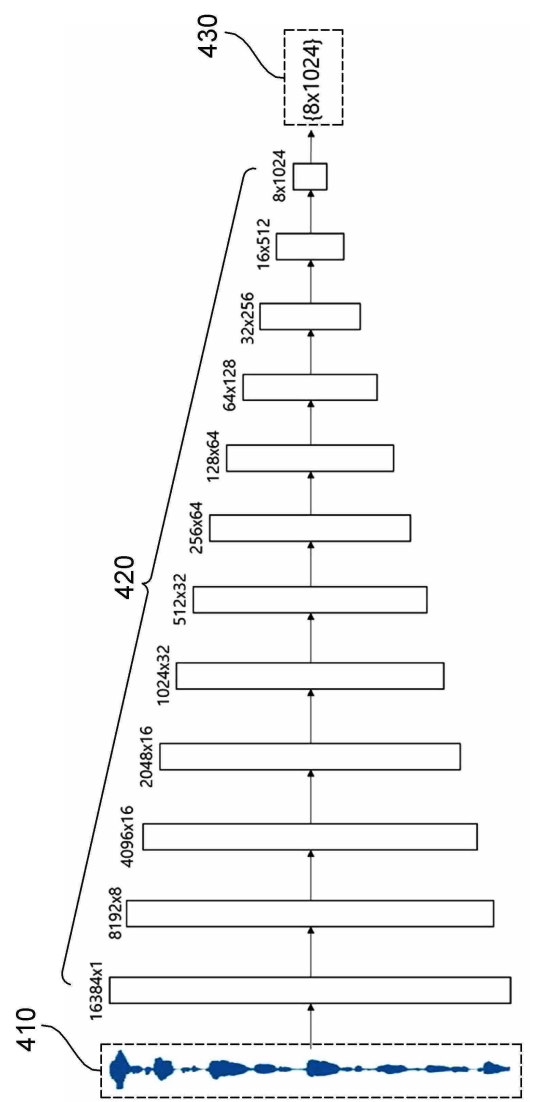
도면2



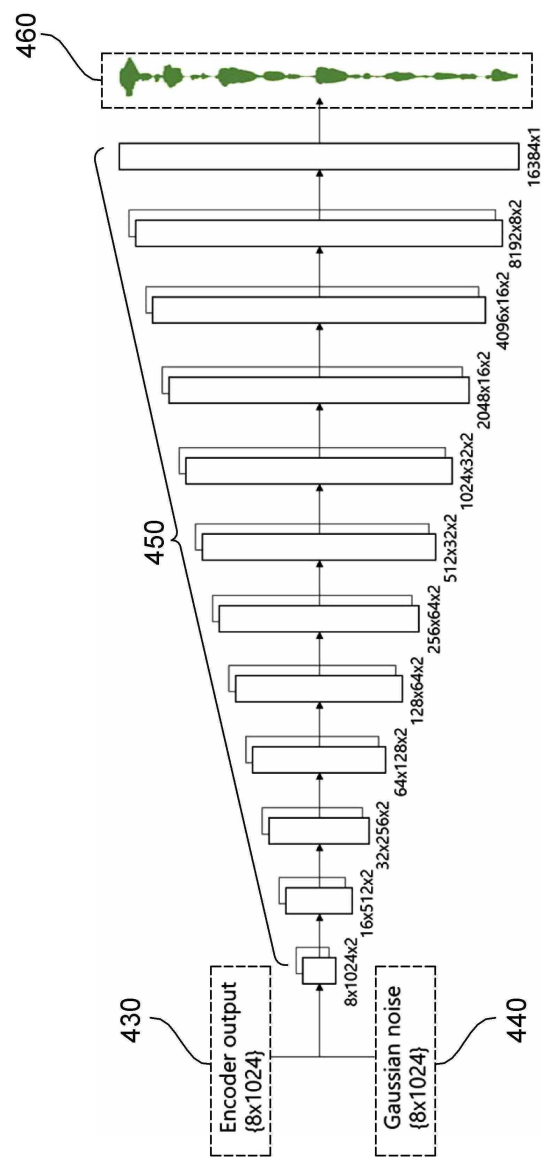
도면3



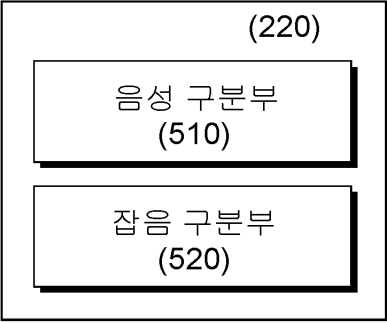
도면4a



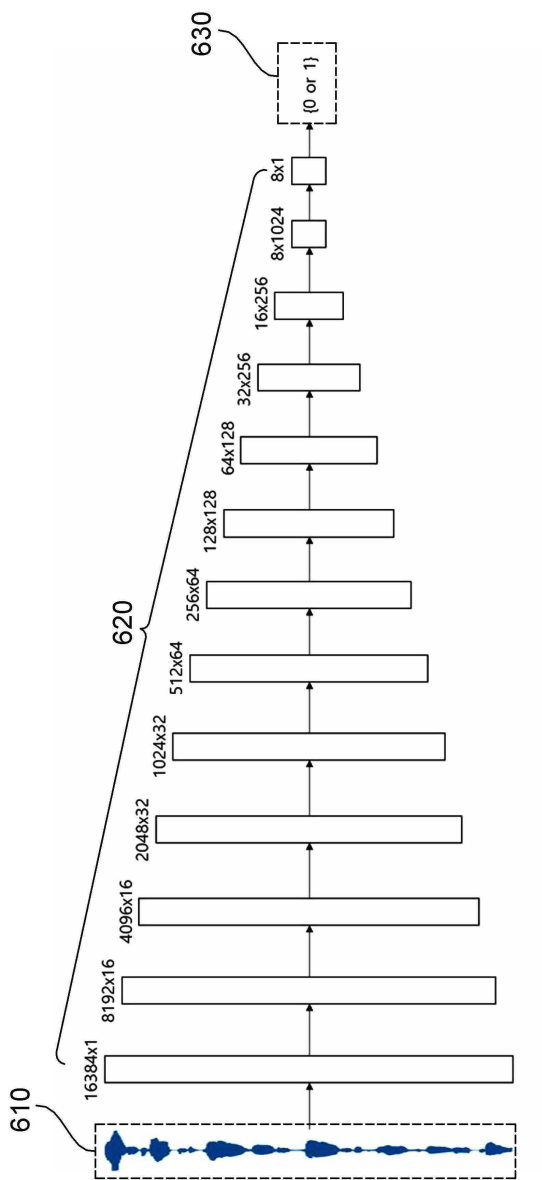
도면4b



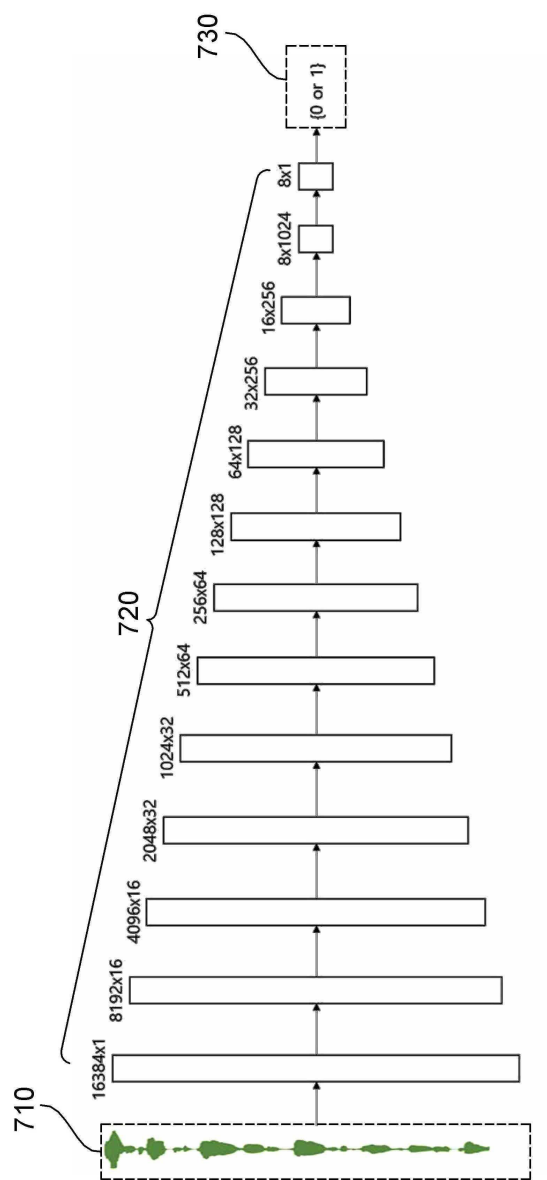
도면5



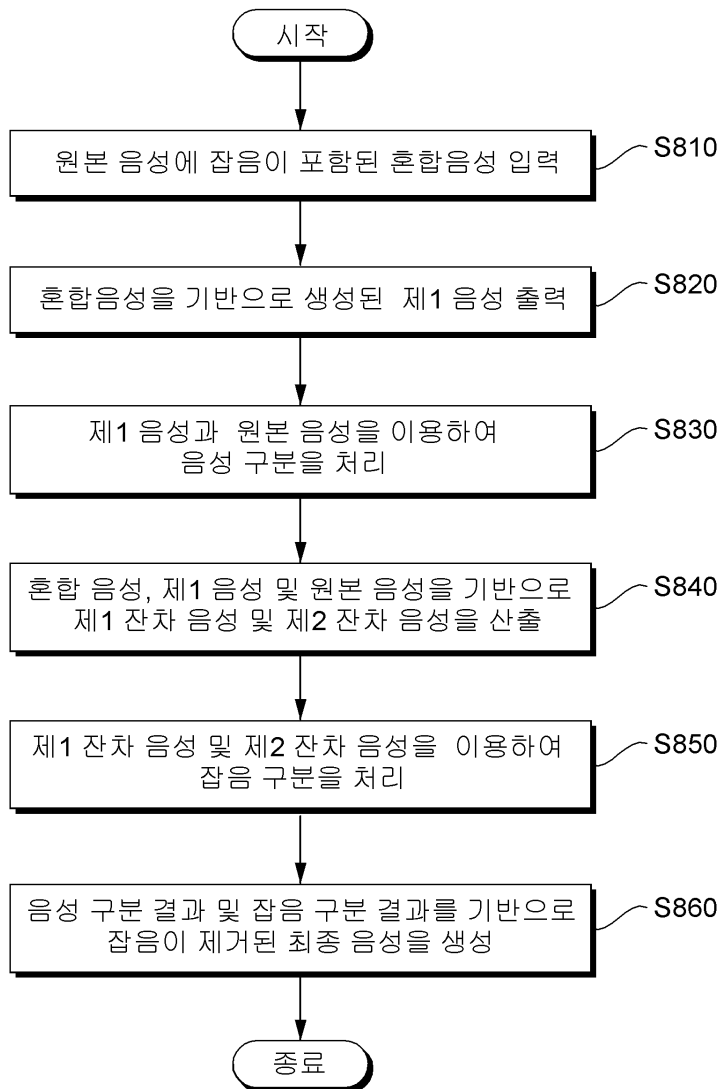
도면6



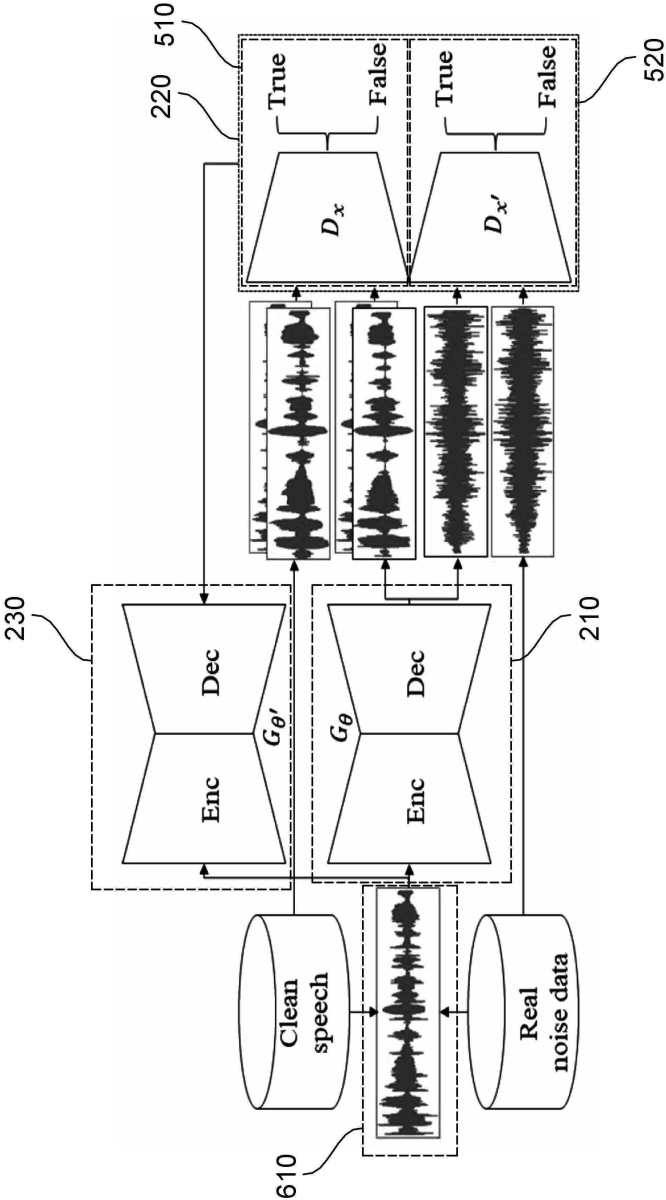
도면7



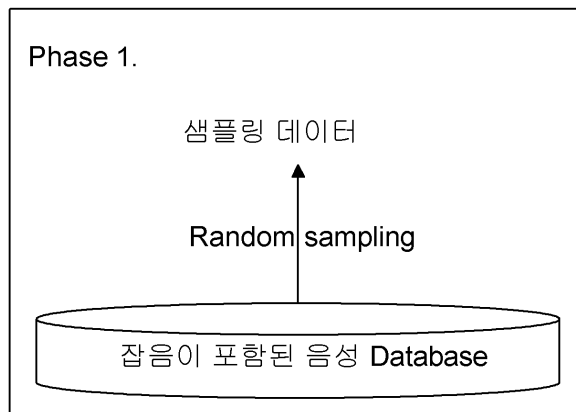
도면8



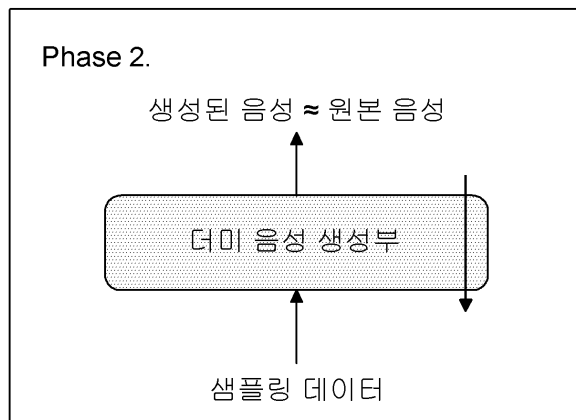
도면9



도면10a

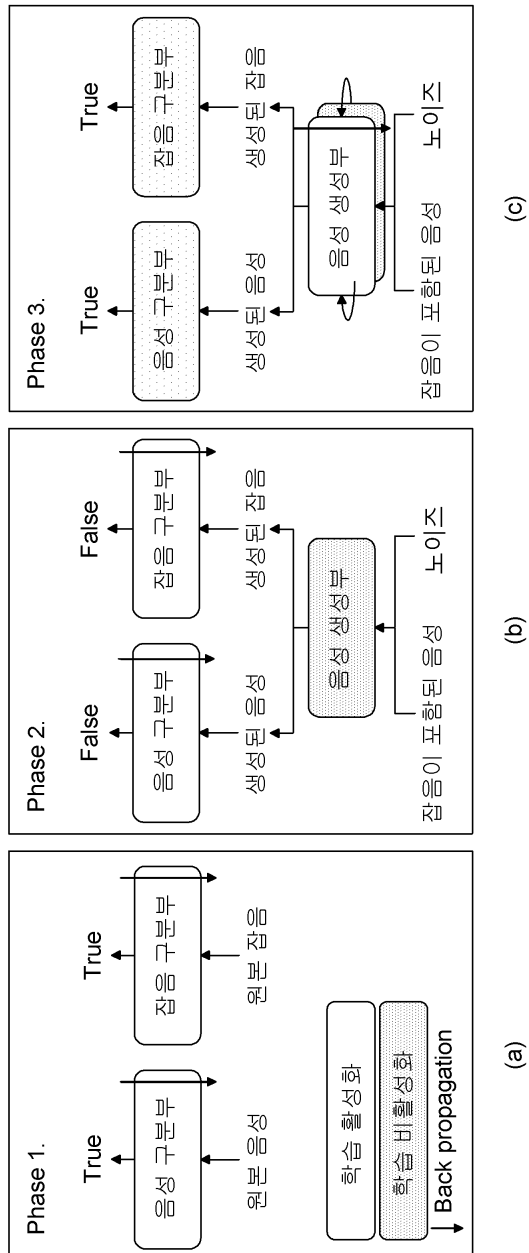


(a)



(b)

도면 10b



도면11

