(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년01월03일
(11) 등록번호 10-2346133
(24) 등록일자 2021년12월28일

(51) 국제특허분류(Int. Cl.)
H04R 3/00 (2006.01) *G01S 3/808* (2006.01)
(52) CPC특허분류
H04R 3/005 (2013.01)
G01S 3/8083 (2013.01)
(21) 출원번호 10-2020-0025548
(22) 출원일자 2020년02월28일
심사청구일자 2020년02월28일
(65) 공개번호 10-2021-0110081
(43) 공개일자 2021년09월07일
(56) 선행기술조사문헌
KR1020120080409 A
KR1020180122171 A
KR1020190041834 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
광주과학기술원
광주광역시 북구 첨단과기로 123 (오룡동)
(72) 발명자
신종원
광주광역시 북구 첨단과기로 123(오룡동) 광주과학기술원 전기전자컴퓨터공학부
박준형
광주광역시 북구 첨단과기로 123(오룡동) 광주과학기술원 전기전자컴퓨터공학부
김민승
광주광역시 북구 첨단과기로 123(오룡동) 광주과학기술원 전기전자컴퓨터공학부
(74) 대리인
김기문

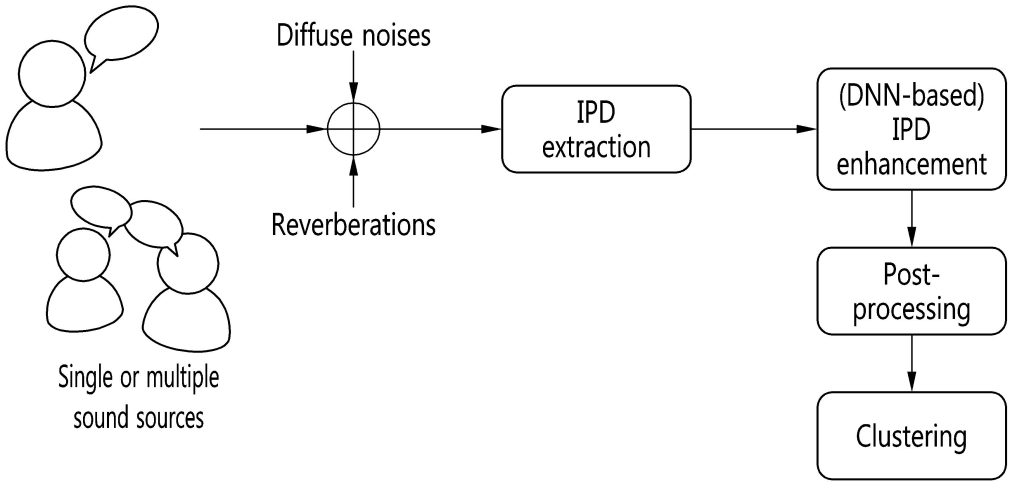
전체 청구항 수 : 총 8 항 심사관 : 강석제

(54) 발명의 명칭 **심층 신경망 기반의 방향각 추정 방법**

(57) 요 약

심층 신경망 기반의 방향각 추정 방법이 개시된다. 본 발명의 실시 예에 따른 심층 신경망 기반의 방향각 추정 방법은, 다 채널 마이크를 통하여, 음원으로부터 생성되고 잡음 및 잔향 중 적어도 하나에 의해 왜곡된 오디오 신호를 수신하는 단계, 상기 오디오 신호의 채널 간 위상 차를 획득하는 단계, 상기 위상 차를 학습된 딥 러닝 모델에 제공하여 깨끗한 오디오 신호의 위상 차를 추정하는 단계, 및, 상기 추정된 위상 차를 이용하여 상기 음원의 방향각을 추정하는 단계를 포함한다.

대 표 도 - 도2



(52) CPC특허분류
H04R 2201/403 (2013.01)
H04R 2430/23 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1415162607(GM12970)
부처명	산업통상자원부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	인공지능융합로봇시스템기술
연구과제명	로봇용 free-running 임베디드 자연어 대화음성인식을 위한 원천 기술 개발
기 여 율	1/1
과제수행기관명	한양대학교산학협력단
연구기간	2019.01.01 ~ 2019.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

다 채널 마이크를 통하여, 음원으로부터 생성되고 잡음 및 잔향 중 적어도 하나에 의해 왜곡된 오디오 신호를 수신하는 단계;

상기 오디오 신호의 채널 간 위상 차를 획득하는 단계;

상기 위상 차를 학습된 딥 러닝 모델에 제공하여, 상기 오디오 신호로부터 잡음과 잔향이 섞이지 않은 깨끗한 오디오 신호의 채널 간 위상 차를 추정하는 단계; 및

상기 추정된 위상 차를 이용하여 상기 음원의 방향각을 추정하는 단계를 포함하는

심층 신경망 기반의 방향각 추정 방법.

청구항 2

제 1항에 있어서,

왜곡된 훈련용 오디오 신호로부터 제1 훈련용 위상 차를 획득하는 단계;

상기 왜곡된 훈련용 오디오 신호에 대응하는 잡음과 잔향이 섞이지 않은 깨끗한 훈련용 오디오 신호로부터 제2 훈련용 위상 차를 획득하는 단계; 및

상기 제1 훈련용 위상 차 및 상기 제2 훈련용 위상 차를 포함하는 훈련용 데이터 셋을 이용하여 심층 신경망을 트레이닝 함으로써 상기 학습된 딥 러닝 모델을 획득하는 단계를 더 포함하는

심층 신경망 기반의 방향각 추정 방법.

청구항 3

제 2항에 있어서,

상기 학습된 딥 러닝 모델은,

회귀(regression) 모델인

심층 신경망 기반의 방향각 추정 방법.

청구항 4

제 3항에 있어서,

상기 학습된 딥 러닝 모델을 획득하는 단계는,

상기 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징에 상기 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징을 레이블 하여 상기 심층 신경망을 트레이닝 하는

심층 신경망 기반의 방향각 추정 방법.

청구항 5

제 4항에 있어서,

상기 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징에 상기 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징을 레이블 하여 상기 심층 신경망을 트레이닝 하는 단계는,

상기 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징 및 상기 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징 간의 비용 함수가 최소가 되도록 상기 심층 신경망을 트레이닝 하는 단계를 포함하고,

상기 비용 함수는, MSE 함수인

심층 신경망 기반의 방향각 추정 방법.

청구항 6

제 1항에 있어서,

상기 위상 차를 학습된 딥 러닝 모델에 제공하여 상기 깨끗한 오디오 신호의 채널 간 위상 차를 추정하는 단계는,

상기 위상 차에 대한 삼각함수 벡터 특징을 상기 학습된 딥 러닝 모델에 입력하고, 상기 딥 러닝 모델에 의해 추정된 상기 깨끗한 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징을 획득하는 단계; 및

상기 깨끗한 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징을 이용하여 상기 깨끗한 오디오 신호의 채널 간 위상 차를 획득하는 단계를 포함하는

심층 신경망 기반의 방향각 추정 방법.

청구항 7

제 1항에 있어서,

상기 음원이 엔드 파이어 방향에 위치할 때의 추정 편차를 보상하기 위하여, 상기 추정된 방향각을 사후처리 하는 단계를 더 포함하는

심층 신경망 기반의 방향각 추정 방법.

청구항 8

제 1항에 있어서,

하나의 프레임 내 복수의 주파수 빈에 각각 대응하는 복수의 방향각을 추정하는 단계; 및

상기 복수의 방향각을 군집화 하고, 군집화의 결과에 기초하여 복수의 음원의 방향각들을 획득하는 단계를 더 포함하는

심층 신경망 기반의 방향각 추정 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은, 두개의 마이크에서 수신된 오디오 신호의 위상 차를 이용하여 깨끗한 오디오 신호의 위상 차를 추정함으로써, 음원의 정확한 방향각을 추정할 수 있는, 심층 신경망 기반의 방향각 추정 방법에 관한 것이다.

배 경 기 술

[0002] 방향각 추정 기법이란, 오디오 신호를 이용하여 하나 이상의 음원(acoustic sound source)의 방향각을 찾는 방법을 의미할 수 있다. 여기서 방향각은 도래각(direction-of-arrival, DoA)이라는 용어와 병행하여 사용될 수 있다.

[0003] 실험실 환경이 아닌 실제 환경에서, 배경잡음(background noise)나 잔향(reverberation)과 같은 간섭신호로 인해 공간정보(spatial information)가 왜곡되는 문제가 발생한다. 따라서 음원의 방향각을 정확히 추정하는 것은 매우 어려운 과제이다.

[0004] 방향각 추정 기법에 대하여 선행 기술 1 (N. Ma et al., “Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments,” IEEE/ACM Trans. Audio, Speech, Lang. Process., 2017)은, DNN 분류모델(DNN classification) 기반의 방향 추정 기법을 제안한다.

[0005] 선행 기술 1은, DNN의 출력 특징(feature)의 차원(dimension)을 방향각의 분류군(class)의 수로 설정하고, DNN으로 획득하는 사후 확률(posterior probability)이 최대가 되는 클래스(class)를 해당 방향각으로 선택하는 방식이다.

[0006] 여기서 DNN의 입력 특징 및 출력 특징은 상호 상관 함수(cross correlation function, CCF)와 방향각들의 클래스(class) 집합일 수 있다.

[0007] 한편 출력 특징(feature)의 차원의 수에 따라 추정하는 방향각의 분해능(resolution)이 결정될 수 있다.

[0008] 예를 들어 -90~+90도 범위의 방향각을 10도 단위로 분류하면 분류군의 수는 19개가 되고, -90~+90도 범위의 방향각을 5도 단위로 분류하면 분류군의 수는 37개, -90~+90도 범위의 방향각을 1도 단위로 분류하면 분류군의 수는 181개가 된다.

[0009] -90~+90도 범위의 방향각을 10도 단위로 분류하는 경우, 추정된 방향각의 분해능(resolution)이 10도로 고정되는 문제가 발생할 수 있다. 또한 분해능 개선을 위해 고차원의 출력 특징을 도입할 경우, 딥러닝에 의한 방향각 추정 성능이 저하될 소지가 있었다. 또한 고차원의 출력 특징을 도입하더라도, 시간-주파수 영역에서의 모든 성분에 대한 방향각 추정은 쉽지가 않다는 문제가 발생할 수 있었다.

발명의 내용

해결하려는 과제

[0010] 본 발명은 상술한 문제점을 해결하기 위한 것으로, 본 발명의 목적은, 두개의 마이크에서 수신된 오디오 신호의 위상 차를 이용하여 깨끗한 오디오 신호의 위상 차를 추정함으로써, 음원의 정확한 방향각을 추정할 수 있는, 심층 신경망 기반의 방향각 추정 방법을 제공하기 위함이다.

과제의 해결 수단

[0011] 본 발명의 실시 예에 다른 심층 신경망 기반의 방향각 추정 방법은, 다 채널 마이크를 통하여, 음원으로부터 생성되고 잡음 및 잔향 중 적어도 하나에 의해 왜곡된 오디오 신호를 수신하는 단계, 상기 오디오 신호의 채널 간 위상 차를 획득하는 단계, 상기 위상 차를 학습된 딥 러닝 모델에 제공하여 깨끗한 오디오 신호의 위상 차를 추정하는 단계, 및, 상기 추정된 위상 차를 이용하여 상기 음원의 방향각을 추정하는 단계를 포함한다.

[0012] 이 경우 왜곡된 훈련용 오디오 신호로부터 획득된 제1 훈련용 위상 차를 획득하는 단계, 상기 왜곡된 훈련용 오디오 신호에 대응하는 깨끗한 훈련용 오디오 신호로부터 획득된 제2 훈련용 위상 차를 획득하는 단계, 및, 상기 제1 훈련용 위상 차 및 상기 제2 훈련용 위상 차를 포함하는 훈련용 데이터 셋을 이용하여 심층 신경망을 트레이닝 함으로써 상기 학습된 딥 러닝 모델을 획득하는 단계를 더 포함할 수 있다.

[0013] 이 경우 상기 학습된 딥 러닝 모델은, 회귀(regression) 모델일 수 있다.

[0014] 이 경우 상기 학습된 딥 러닝 모델을 획득하는 단계는, 상기 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징에 상기 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징을 레이블 하여 상기 심층 신경망을 트레이닝 할 수 있다,

[0015] 이 경우 상기 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징에 상기 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징을 레이블 하여 상기 심층 신경망을 트레이닝 하는 단계는, 상기 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징 및 상기 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징 간의 비용 함수가 최소가 되도록 상기 심층 신경망을 트레이닝 하는 단계를 포함하고, 상기 비용 함수는, MSE 함수일 수 있다.

[0016] 한편 상기 위상 차를 학습된 딥 러닝 모델에 제공하여 깨끗한 오디오 신호의 위상 차를 추정하는 단계는, 상기 위상 차에 대한 삼각함수 벡터 특징을 상기 학습된 딥 러닝 모델에 입력하고, 상기 딥 러닝 모델에 의해 추정된 상기 깨끗한 오디오 신호의 위상 차에 대한 삼각함수 벡터 특징을 획득하는 단계, 및, 상기 깨끗한 오디오 신호의 위상 차에 대한 삼각함수 벡터 특징을 이용하여 상기 깨끗한 오디오 신호의 위상 차를 획득하는 단계를 포함할 수 있다.

[0017] 한편 상기 음원이 엔드 파이어 방향에 위치할 때의 추정 편차를 보상하기 위하여, 상기 추정된 방향각을 사후처리 하는 단계를 더 포함할 수 있다.

[0018] 한편 하나의 프레임의 복수의 주파수 빈에 각각 대응하는 복수의 방향각을 추정하는 단계, 및, 상기 복수의 방향각을 군집화 하고, 군집화의 결과에 기초하여 복수의 음원의 방향각들을 획득하는 단계를 더 포함할 수 있다.

도면의 간단한 설명

[0019] 도 1은 방향각 추정 장치를 설명하기 위한 블록도이다.

도 2는 심층 신경망 기반의 방향각 추정 방법의 개요를 설명하기 위한 도면이다.

도 3은 본 발명의 실시 예에 따른, 학습된 딥 러닝 모델을 획득하는 방법을 설명하기 위한 도면이다.

도 4는 본 발명의 실시 예에 따른, 학습된 딥 러닝 모델을 이용하여 음원의 방향각을 정확하게 추정하는 방법을 설명하기 위한 도면이다.

도 5는 본 발명의 실시 예에 따른, 2채널 마이크 및 2채널 마이크에서 각각 수신된 오디오 신호를 도시한 도면이다.

도 7은 단일 음원이 40도 각도에 있을 때 채널간 위상 차가 향상된 결과를 도시한 도면이다.

도 8은 Babble 노이즈가 5dB의 SNR로 존재하는 경우, 다양한 음원 위치에서의 실제 음원의 방향 각과 추정된 음원의 방향각을 비교한 도면이다.

도 9는 딥 러닝 모델을 이용하여 방향각을 직접 추정하는 방식, 방향각의 정현파 함수를 추정하는 방식, 채널간 위상 차의 정현파 함수를 추정하는 방식의 실험 결과를 도시한 도면이다.

도 10은 또 다른 테스트에서의 조건을 설명한 도면이다.

도 11은 도 10의 조건에서, 노이즈를 조절해 가면서, 세가지 방향각 추정 기법을 사용하여 실험한 결과를 도시한 도면이다.

도 12는 도 10의 조건에서, 잔향 시간(RT60)을 조절해 가면서, 세가지 방향각 추정 기법을 사용하여 실험한 결과를 도시한 도면이다.

도 13은 실험 결과를 노이즈의 종류 별로 세분화 한 결과를 도시한 도면이고, 도 14는 실험 결과를 두개의 잔향 시간으로 세분화 한 결과를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 이하, 첨부된 도면을 참조하여 본 명세서에 개시된 실시 예를 상세히 설명하되, 도면 부호에 관계없이 동일하거나 유사한 구성요소는 동일한 참조 번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다. 또한, 본 명세서에 개시된 실시 예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 명세서에 개시된 실시 예의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다. 또한, 첨부된 도면은 본 명세서에 개시된 실시 예를 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 명세서에 개시된 기술적 사상이 제한되지 않으며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0021] 제1, 제2 등과 같이 서수를 포함하는 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되지는 않는다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0022] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0023] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함한다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0024] 도 1은 방향각 추정 장치를 설명하기 위한 블록도이다.

[0025] 방향각 추정 장치는, 심층 신경망 기반의 방향각 추정 방법을 수행하는 장치일 수 있다.

[0026] 방향각 추정 방법이란, 오디오 신호를 이용하여 하나 이상의 음원(acoustic sound source)의 방향각을 찾는 방법을 의미할 수 있다.

[0027] 여기서 오디오 신호란, 음성 신호 및 기타 가청 범위 내의 음향 신호를 포함할 수 있다.

[0028] 또한 방향각은, 다 채널 마이크를 기준으로 한 음원의 위치를 나타내는 것으로, 도래각(direction-of-arrival, DoA)이라는 용어와 병행하여 사용될 수 있다.

[0029] 본 발명의 실시 예에 따른 방향각 추정 장치(100)는, 출력부(120), 수신부(110), 제어부(130) 및 메모리(140)를 포함할 수 있다.

[0030] 수신부(110)는 오디오 신호를 수신할 수 있다.

[0031] 구체적으로 수신부(110)는 다 채널 마이크를 포함할 수 있으며, 다 채널 마이크는 외부로부터 오디오 신호를 수신할 수 있다. 여기서 다 채널 마이크는 둘 이상의 마이크를 포함할 수 있다.

[0032] 메모리(140)는 방향각 추정 장치(100)의 다양한 기능을 지원하는 데이터를 저장할 수 있다.

[0033] 출력부(120)는 추정된 방향각을 출력할 수 있다. 구체적으로 출력부(120)는 디스플레이 및 스피커 중 적어도 하나를 포함하고, 제어부(130)에 의하여 추정된 방향각을 디스플레이 하거나 음향 신호로 출력할 수 있다.

[0034] 한편 제어부(130)는 방향각 추정 장치(100)의 전반적인 동작을 제어할 수 있다.

[0035] 또한 음원으로부터 생성된 오디오 신호가 다 채널 마이크를 통하여 수신되면, 제어부(130)는 수신된 오디오 신호를 이용하여 음원의 방향각을 추정할 수 있다.

[0036] 도 2는 심층 신경망 기반의 방향각 추정 방법의 개요를 설명하기 위한 도면이다.

[0037] 방향각 추정 장치 주변에는 하나 이상의 음원(sound source)이 존재하고, 하나 이상의 음원 각각에 의해 오디오 신호가 생성될 수 있다.

[0038] 한편 음원에 의해 생성된 오디오 신호는, 주변의 잡음(noise) 및 잔향(reverberation)에 의해 왜곡될 수 있다. 따라서 방향각 추정 장치의 다 채널 마이크에는, 잡음(noise) 및 잔향(reverberation) 중 적어도 하나에 의해 왜곡된 오디오 신호가 수신되게 된다.

[0039] 한편 방향각 추정 장치는, 왜곡된 오디오 신호를 이용하여 공간 정보를 획득할 수 있다.

[0040] 여기서 공간 정보는, 음원의 방향각 추정에 활용되는 정보로써, 채널간 시간차(interchannel time difference, ITD), 채널간 위상차(interchannel phase difference, IPD) 및 채널간 레벨차(interchannel level difference, ILD)를 포함할 수 있다.

[0041] 채널간 시간차(interchannel time difference, ITD)는 두 개 이상의 마이크로 획득한 오디오 신호 들 사이의 시간 차를 의미할 수 있다.

[0042] 또한 채널간 위상차(interchannel phase difference, IPD)는, 두 개 이상의 마이크로 획득한 오디오 신호들 사이의 위상 차를 의미할 수 있다.

[0043] 또한 채널간 레벨차(interchannel level difference, ILD)는, 두 개 이상의 마이크로 획득한 오디오 신호들 사이의 레벨 차를 의미할 수 있다.

[0044] 본 발명에서는 공간 정보 중 채널간 위상차(interchannel phase difference, IPD)를 향상시키는 방법에 대하여 설명한다. 다만 이에 한정되지 않으며, 본 발명은 채널간 시간차(interchannel time difference, ITD)나 채널간 레벨차(interchannel level difference, ILD)를 향상시키는 방법에도 적용될 수 있다.

[0045] 방향각 추정 장치는, 왜곡된 오디오 신호를 이용하여 채널 간 위상 차, 즉 두 개의 마이크에서 각각 수신된 오디오 신호들 간의 위상 차를 획득할 수 있다. 다만 이와 같이 획득한 채널 간 위상 차는 왜곡된 오디오 신호에 기반하여 획득된 것이기 때문에, 채널 간 위상 차 역시 왜곡된 상태일 수 있다.

[0046] 이 경우 방향각 추정 장치는 획득된 채널 간 위상 차를 학습된 딥 러닝 모델에 제공하여 깨끗한 오디오 신호의 채널 간 위상 차를 추정할 수 있다. 여기서 깨끗한 오디오 신호란, 왜곡된 오디오 신호로부터 잡음과 잔향을 줄인 신호를 의미할 수 있다.

[0047] 이 경우 방향각 추정 장치는 추정된 위상차를 이용하여 음원의 방향각을 추정할 수 있다. 또한 방향각 추정 장

치는 추정된 방향각을 사후 처리하여, 방향각 추정의 성능을 향상시킬 수도 있다.

[0048] 또한 음원이 복수개인 경우, 방향각 추정 장치는 복수의 주파수에 각각 대응하는 복수의 방향각을 군집화 하여 복수의 음원 각각에 대한 방향각을 추정할 수 있다.

[0049] 한편 본 발명에서는 학습된 딥 러닝 모델을 이용하여 깨끗한 위상 차를 추정하게 된다. 따라서 학습된 딥 러닝 모델을 획득하는 방법에 대하여 설명한다.

[0050] 도 3은 본 발명의 실시 예에 따른, 학습된 딥 러닝 모델을 획득하는 방법을 설명하기 위한 도면이다.

[0051] 심층 신경망을 트레이닝 하여 학습된 딥 러닝 모델을 획득하는 장치를 학습 장치라고 명칭 하도록 한다. 여기서 학습 장치는, 도 1에서 설명한 방향각 추정 장치(100)의 구성을 포함할 수 있다.

[0052] 다만 심층 신경망을 트레이닝 하여 학습된 딥 러닝 모델을 획득하는 과정은 방향각 추정 장치(100) 에 의해서도 수행될 수 있다.

[0053] 학습 장치는 잡음 및 잔향이 섞이지 않은 깨끗한 훈련용 오디오 신호와, 잡음 및 잔향 중 적어도 하나를 포함하는 왜곡된 훈련용 오디오 신호를 획득할 수 있다.

[0054] 여기서 왜곡된 훈련용 오디오 신호는 깨끗한 훈련용 오디오 신호에 대응할 수 있다. 구체적으로 왜곡된 훈련용 오디오 신호와 깨끗한 훈련용 오디오 신호는, 동일한 조건(동일한 환경, 동일한 음원의 위치 등)에서 동일한 음원으로부터 획득되는 것일 수 있다.

[0055] 예를 들어 훈련용 음원이 깨끗한 훈련용 오디오 신호를 출력하는 경우, 학습 장치는 깨끗한 훈련용 오디오 신호와, 깨끗한 훈련용 오디오 신호에 잡음 및 잔향 중 적어도 하나가 첨가된 왜곡된 오디오 신호를 획득할 수 있다.

[0056] 한편 학습 장치는, 왜곡된 훈련용 오디오 신호에 다운 믹싱을 적용할 수 있다(S305).

[0057] 구체적으로 왜곡된 훈련용 오디오 신호가 3채널 이상의 마이크에 의해 수신된 경우, 학습 장치는 3채널 이상의 마이크에 의해 수신된 오디오 신호를 2채널 마이크에 의해 수신된 오디오 신호로 변환할 수 있다.

[0058] 한편 학습 장치는, 왜곡된 훈련용 오디오 신호의 채널 간 위상 차(제1 훈련용 위상 차)를 획득할 수 있다(S310).

[0059] 구체적으로, 학습 장치는 왜곡된 훈련용 오디오 신호에 단구간 푸리에 변환(short-time Fourier transform, STFT)을 적용하여 왜곡된 훈련용 오디오 신호의 복소 스펙트럼(complex spectrum)을 획득하고, 왜곡된 훈련용 오디오 신호의 복소 스펙트럼에 대한 위상 정보를 추출할 수 있다.

[0060] 그리고 학습 장치는, 2채널 마이크 중 제1 마이크에서 수신된 왜곡된 훈련용 오디오 신호와 2채널 마이크 중 제2 마이크에서 수신된 왜곡된 오디오 신호 간의 위상 차인 제1 훈련용 위상 차를 획득할 수 있다.

[0061] 한편 학습 장치는 왜곡된 오디오 신호의 채널 간 위상 차(제1 훈련용 위상 차)를 삼각 함수 벡터 특징으로 변환할 수 있다(S315).

[0062] 이와 관련하여, 본 발명의 학습된 딥 러닝 모델은 회귀(regression) 모델일 수 있다.

[0063] 즉 선행 기술 1에서는 출력 특징의 차원이 클래스의 수에 한정되어 방향각의 분해능이 떨어지게 된다. 본 발명에서는 이러한 문제를 해결하기 위하여 딥 러닝 모델을 회귀(regression) 모델로 트레이닝 할 수 있다.

[0064] 다만 채널 간 위상 차(제1 훈련용 위상 차)는 ???부터 ??까지 구간의 값을 가지는 불연속성을 가지고 있다.

[0065] 따라서 채널 간 위상 차(제1 훈련용 위상 차)로는, 딥 러닝 회귀(regression) 모델의 비용 함수로 활용되는 평균제곱오차(mean square error, MSE) 함수를 활용할 수 없다.

[0066] 따라서 평균제곱오차(mean square error, MSE) 함수와 같은 비용 함수로도 신경망의 훈련이 가능하도록 하기 위해, 채널 간 위상 차(제1 훈련용 위상 차)에 삼각 함수를 취하는 방식으로 불연속성을 해결할 수 있다. 이와 관련해서는 이후에 더욱 자세히 설명하도록 한다.

[0067] 한편 학습 장치는 깨끗한 훈련용 오디오 신호에 대하여, S305, S310, S315와 동일한 처리를 수행할 수 있다.

[0068] 구체적으로 학습 장치는, 깨끗한 훈련용 오디오 신호에 다운 믹싱을 적용할 수 있다(S320).

[0069] 또한 학습 장치는, 깨끗한 훈련용 오디오 신호의 채널 간 위상 차(제2 훈련용 위상 차)를 획득할 수 있다

(S325).

[0070] 구체적으로, 학습 장치는 깨끗한 훈련용 오디오 신호에 단구간 푸리에 변환(short-time Fourier transform, STFT)을 적용하여 깨끗한 훈련용 오디오 신호의 복소 스펙트럼(complex spectrum)을 획득하고, 깨끗한 훈련용 오디오 신호의 복소 스펙트럼에 대한 위상 정보를 추출할 수 있다.

[0071] 그리고 학습 장치는, 2채널 마이크 중 제1 마이크에서 수신된 깨끗한 훈련용 오디오 신호와 2채널 마이크 중 제2 마이크에서 수신된 깨끗한 훈련용 오디오 신호 간의 위상 차인 제1 훈련용 위상 차를 획득할 수 있다.

[0072] 한편 학습 장치는 깨끗한 오디오 신호의 채널 간 위상 차(제2 훈련용 위상 차)를 삼각 함수 벡터 특징으로 변환할 수 있다(S330).

[0073] 구체적으로 채널 간 위상 차(제2 훈련용 위상 차)의 불연속성을 해결하기 위하여, 학습 장치는 채널 간 위상 차(제2 훈련용 위상 차)에 삼각 함수를 취할 수 있다.

[0074] 한편 학습 장치는 제1 훈련용 위상 차 및 제2 훈련용 위상 차를 포함하는 훈련용 데이터 셋을 이용하여 심층 신경망을 트레이닝 함으로써, 학습된 딥 러닝 모델을 생성할 수 있다(S335, S340).

[0075] 하나의 훈련용 데이터 셋을 구성하는 제1 훈련용 위상 차 및 제2 훈련용 위상 차는, 서로 대응할 수 있다. 구체적으로 하나의 훈련용 데이터 셋을 구성하는 제1 훈련용 위상 차 및 제2 훈련용 위상 차는 동일한 조건(동일한 환경, 동일한 음원의 위치 등)에서 동일한 음원으로부터 획득되는 것일 수 있다.

[0076] 한편 심층 신경망은 지도 학습(supervised learning) 방식, 그 중에서도 회귀(regression) 분석 방식으로 트레이닝 할 수 있다.

[0077] 구체적으로 학습 장치는, 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징에 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징을 레이블 하여 심층 신경망을 트레이닝 할 수 있다.

[0078] 이 경우 왜곡된 훈련용 오디오 신호의 제1 훈련용 위상 차(구체적으로 제1 훈련용 위상차에 대한 삼각 함수 벡터 특징)이 심층 신경망의 입력으로 활용되고, 깨끗한 훈련용 오디오 신호의 제2 훈련용 위상 차(구체적으로 제2 훈련용 위상차에 대한 삼각 함수 벡터 특징)이 심층 신경망의 출력으로 활용될 수 있다. 또한 입력과 출력 간의 차이는 평균제곱오차(mean square error, MSE) 함수를 활용하여 역 전파될 수 있으며, 학습 장치는 제1 훈련용 위상 차에 대한 삼각함수 벡터 특징 및 제2 훈련용 위상 차에 대한 삼각함수 벡터 특징 간의 비용 함수가 최소가 되도록 심층 신경망을 트레이닝 할 수 있다.

[0079] 그리고 복수의 훈련용 데이터 셋을 이용하여 심층 신경망이 트레이닝 됨에 따라 최적화된 파라미터를 가지는 학습된 딥 러닝 모델이 생성될 수 있다.

[0080] 한편 학습된 딥 러닝 모델은 방향각 추정 장치(100)에 탑재될 수 있다. 이 경우 학습된 딥 러닝 모델을 구성하는 하나 이상의 명령어는, 방향각 추정 장치(100)의 메모리(140)에 저장될 수 있다.

[0081] 도 4는 본 발명의 실시 예에 따른, 학습된 딥 러닝 모델을 이용하여 음원의 방향각을 정확하게 추정하는 방법을 설명하기 위한 도면이다.

[0082] 방향각 추정 장치(100)의 제어부(130)는 다 채널 마이크를 통하여, 음원으로부터 생성되고 잡음 및 잔향 중 적어도 하나에 의해 왜곡된 오디오 신호를 수신할 수 있다(S405).

[0083] 그리고 제어부(130)는 왜곡된 오디오 신호에 다운 믹싱을 적용할 수 있다(S410).

[0084] 구체적으로 왜곡된 훈련용 오디오 신호가 3채널 이상의 마이크에 의해 수신된 경우, 제어부(130)는 3채널 이상의 마이크에 의해 수신된 오디오 신호를 2채널 마이크에 의해 수신된 오디오 신호로 변환할 수 있다.

[0085] 한편 왜곡된 오디오 신호가 2 채널 마이크에 의해 수신된 경우, S405는 생략될 수 있다.

[0086] 한편 제어부(130)는, 왜곡된 오디오 신호의 채널 간 위상 차를 획득할 수 있다(S415).

[0087] 이와 관련하여 도 5를 참고하여 설명한다.

[0088] 도 5는 본 발명의 실시 예에 따른, 2채널 마이크 및 2채널 마이크에서 각각 수신된 오디오 신호를 도시한 도면이다.

[0089] 구체적으로 제어부(130)는, 왜곡된 오디오 신호에 단구간 푸리에 변환(short-time Fourier transform, STFT)을

적용하여 왜곡된 오디오 신호의 복소 스펙트럼(complex spectrum)을 획득하고, 왜곡된 오디오 신호의 복소 스펙트럼에 대한 위상 정보를 추출할 수 있다.

[0090] 그리고 제어부(130)는, 2채널 마이크 중 제1 마이크(mic 1)에서 수신된 왜곡된 오디오 신호(510)와 2채널 마이크 중 제2 마이크(mic 2)에서 수신된 왜곡된 오디오 신호(520) 간의 위상 차인 채널 간 위상 차($\Delta\phi(k,m)$)를 획득할 수 있다.

[0091] 여기서 m은 m번째 프레임을 의미할 수 있으며, k는 k번째 주파수 빈(frequency bin)을 의미할 수 있다. 따라서 $\Delta\phi(k,m)$는 m번째 프레임 내 k번째 주파수 빈(frequency bin)에서의 채널 간 위상 차($\Delta\phi(k,m)$)를 의미할 수 있다.

[0092] 한편 채널 간 위상 차($\Delta\phi(k,m)$)를 이용하여 음원의 방향각을 추정할 수 있다. 채널 간 위상 차($\Delta\phi(k,m)$)를 이용하여 음원의 방향각을 추정하는 방법은 아래와 같은 수학식으로 나타낼 수 있다.

수학식 1

[0093]
$$\theta(k,m) = \arcsin\left(\frac{c \cdot \Delta\phi(k,m)}{2\pi f d}\right)$$

[0094] 여기서 $\theta(k,m)$은 음원의 방향 각을 의미할 수 있다. 또한 $\Delta\phi(k,m)$는 채널 간 위상 차, c는 음속(343 m/s), d는 마이크 간 거리, k번째 빈(bin)에 해당하는 주파수를 의미할 수 있다.

[0095] 즉 왜곡된 오디오 신호를 이용하여도 음원의 방향각을 산출할 수 있다. 다만 왜곡된 오디오 신호를 이용하는 경우, 주변 소음, 잔향, 간섭 등으로 인하여 방향각 추정 성능이 저하될 수 있다.

[0096] 따라서 본 발명에서는, 학습된 딥 러닝 모델을 이용하여 채널 간 위상 차($\Delta\phi(k,m)$)를 직접적으로 향상시키는 방법을 취한다. 이 경우 잡음이나 잔향에 의해 왜곡된 채널 간 위상 차($\Delta\phi(k,m)$)를 이용하는 것이 아니라, 깨끗한 채널간 위상 차를 이용하는 것이기 때문에, 음원의 방향각을 보다 정확하게 추정할 수 있다.

[0097] 이를 위하여 제어부(130)는, 위상 차를 학습된 딥 러닝 모델에 제공하여 깨끗한 오디오 신호의 위상 차를 추정할 수 있다.

[0098] 다시 도 4로 돌아가서, 깨끗한 오디오 신호의 위상 차를 추정하기 위하여, 제어부(130)는 왜곡된 오디오 신호의 채널 간 위상 차를 삼각 함수 벡터 특징으로 변환할 수 있다(S420). 또한 제어부(130)는 채널 간 위상 차에 대한 삼각 함수 벡터 특징을 학습된 딥 러닝 모델에 제공하여, 딥 러닝 모델에 의해 추정된 깨끗한 오디오 신호의 위상 차에 대한 삼각함수 벡터 특징을 획득할 수 있다(S425).

[0099] 이와 관련하여 도 6을 참고하여 구체적으로 설명한다.

[0100] 도 6은 깨끗한 오디오 신호의 채널 간 위상차를 획득하는 방법을 설명하기 위한 도면이다.

[0101] 본 발명에서는 채널 간 위상 차에 대한 주 종류의 삼각함수 값(sin, cos)을 모두 이용하는 데, 이는 이 두 종류의 삼각함수가 모두 확보되어야만 깨끗한 위상 차를 정확하게 복원할 수 있기 때문이다.

[0102] 그리고 깨끗한 오디오 신호의 위상 차를 추정하기 위하여, 제어부(130)는 왜곡된 오디오 신호의 채널 간 위상

차($\Delta\phi_y(k,m)$)를 삼각 함수 벡터 특징으로 변환할 수 있다

[0103] 여기서 왜곡된 오디오 신호의 채널 간 위상 차($\Delta\phi_y(k,m)$)에 대한 삼각 함수 벡터 특징(즉 학습된 딥 러닝 모델의 입력 벡터)은 $(\sin\Delta\phi_y(1,m),\ldots,\sin\Delta\phi_y(\frac{K}{2}-1,m),\cos\Delta\phi_y(1,m),\ldots,\cos\Delta\phi_y(\frac{K}{2}-1,m))$로 표현될 수 있다.

[0104] 그리고 제어부(130)는 왜곡된 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징을 학습된 딥 러닝 모델에 입력할 수 있다.

[0105] 이 경우 학습된 딥 러닝 모델은, 왜곡된 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징에 기반하여 깨끗한 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징을 출력할 수 있다. 여기서 깨끗한 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징은 $(\sin\Delta\phi_s(1,m),\ldots,\sin\Delta\phi_s(\frac{K}{2}-1,m),\cos\Delta\phi_s(1,m),\ldots,\cos\Delta\phi_s(\frac{K}{2}-1,m))$로 표현될 수 있다.

[0106] 한편 제어부(130)는 깨끗한 오디오 신호의 채널 간 위상 차에 대한 삼각함수 벡터 특징을 이용하여, 깨끗한 오디오 신호의 채널 간 위상 차($\widehat{\Delta\phi_s}(k,m)$)를 획득할 수 있다. 이는 아래와 같은 수학식으로 나타낼 수 있다.

수학식 2

[0107] $$\widehat{\Delta\phi_s}(k,m)=\arctan 2(\widehat{\sin\Delta\phi_s}(k,m),\widehat{\cos\Delta\phi_s}(k,m))$$

[0108] 즉 제어부(130)는 학습된 딥 러닝 모델이 추정한 삼각함수 벡터 특징에 삼각함수 역변환을 수행하여, 깨끗한 오디오 신호의 채널 간 위상 차($\widehat{\Delta\phi_s}(k,m)$)를 획득할 수 있다.

[0109] 도 7은 단일 음원이 40도 각도에 있을 때 채널간 위상 차가 향상된 결과를 도시한 도면이다.

[0110] 도 7a는 본래 깨끗한 오디오 신호의 채널 간 위상 차, 도 7b는 Babble 노이즈가 5dB의 SNR로 발생한 왜곡된 오디오 신호의 채널 간 위상 차, 도 7c는 도 7b의 채널 간 위상 차를 학습된 딥 러닝 모델을 이용하여 향상시킨 채널 간 위상 차(즉 딥 러닝 모델에 의하여 추정된 채널 간 위상 차), 도 7d는 RT60 = 0.4 초의 잔향을 포함하는 오디오 신호의 채널 간 위상 차, 도 7e는 도 7d의 채널 간 위상 차를 학습된 딥 러닝 모델을 이용하여 향상시킨 채널 간 위상 차(즉 딥 러닝 모델에 의하여 추정된 채널 간 위상 차)이다.

[0111] 도 7을 참고하면, 도 7c는 도 7b에 비하여, 도 7e는 도 7d에 비하여 상당히 향상된 결과를 나타내는 것을 알 수 있다.

[0112] 또한 도 7c와 도 7e를 도 7a와 비교해보면, 딥 러닝 모델에 의해 추정된 위상 차(도 7c, 도 7e)가, 본래 깨끗한 오디오 신호의 채널 간 위상 차(도 7a)와 가깝게 복원된 것을 알 수 있다.

[0113] 다시 도 4로 돌아가서, 제어부(130)는 깨끗한 오디오 신호의 채널 간 위상 차($\widehat{\Delta\phi_s}(k,m)$)를 이용하여 음원의 방향각($\hat{\theta}(k,m)$)을 추정할 수 있다(S460).

[0114] 구체적으로 제어부(130)는 깨끗한 오디오 신호의 채널 간 위상 차($\widehat{\Delta\phi_s}(k,m)$)를 수학식 1에 대입하여, 음원의 방향각($\hat{\theta}(k,m)$)을 추정할 수 있다.

[0115] 도 8은 Babble 노이즈가 5dB의 SNR로 존재하는 경우, 다양한 음원 위치에서의 실제 음원의 방향 각과 추정된 음

원의 방향각을 비교한 도면이다.

[0116] 도 8을 참고하면, 대부분의 음원의 위치에서, 실제 음원의 방향 각과 추정된 음원의 방향각($\hat{\theta}$)이 일치하는 것을 알 수 있다.

[0117] 다시 도 4로 돌아가서, 한편 제어부(130)는 음원이 엔드 파이어 방향에 위치할 때의 추정 편차를 보상하기 위하여, 추정된 방향각을 사후 처리 할 수 있다(S435).

[0118] 구체적으로, 채널 간 위상 차를 이용하는 방향각 추정 기법의 경우, 음원이 음원이 엔드 파이어 방향(end-fire direction) (-90도 또는 +90도)에 위치할 때 방향각 추정에 따른 분해능 저하가 심하게 발생할 수 있다.

[0119] 도 4를 다시 참고하면, 실제 음원의 방향 각과 추정된 음원의 방향각($\hat{\theta}$)은 대부분의 음원 위치에서 일치하나, 음원이 -90도 또는 +90도에 위치하는 경우 실제 음원의 방향 각과 추정된 음원의 방향각($\hat{\theta}$) 사이에 편차가 발생하는 것을 알 수 있다.

[0120] 이것은 수학식 1을 이용한 방향각의 추정이 arcsin 함수를 통해 수행되기 때문에, 1.0보다 -1.0보다 작을 때 정보를 버리거나 arcsin 함수의 인수를 잘라 없애기(truncate) 때문이다. 따라서 추정된 방향각은 90도보다 크거나 -90도보다 작을 수 없기 때문에, 음원의 실제 방향 각이 -90도 또는 +90도에 가까울 때, 방향각들의 평균에는 편차가 발생하게 된다.

[0121] 따라서 이러한 편차를 보정하기 위하여, 제어부(160)는 추정된 방향각을 사후처리 할 수 있다. 이러한 사후 처리는 아래와 같은 수학식으로 나타낼 수 있다.

수학식 3

[0122] $$\hat{\theta}_{adj}(k,m) = \min\left(\max\left(\tilde{\theta}(k,m), -90^{\circ}\right), 90^{\circ}\right)$$

수학식 4

[0123] $$\tilde{\theta}(k,m) = \hat{\theta}(k,m) + a \cdot \mathrm{sgn}(\hat{\theta}(k,m)) \cdot (\hat{\theta}(k,m))^4$$

[0124] 여기서 $\hat{\theta}(k,m)$, $\hat{\theta}_{adj}(k,m)$, sgn (x) 및 a는 각각 추정된 방향 각, 보정된 방향 각, x의 부호(sign) 및 양의 상수를 의미할 수 있다.

[0125] 도 8은 Babble 노이즈가 5dB의 SNR로 존재하는 경우, 다양한 음원 위치에서의 음원의 실제 방향 각과 음원의 보정된 방향각을 비교한 도면이다.

[0126] 도 8을 참고하면, 보정된 방향 각($\hat{\theta}_{adj}$)은 음원의 추정된 방향각($\hat{\theta}$)에 비하여 오차가 보상된 것을 알 수 있다.

[0127] 또한 음원의 모든 위치에서, 보정된 방향 각($\hat{\theta}_{adj}$)과 음원의 실제 방향 각이 일치하는 것을 알 수 있다.

[0128] 다시 도 4로 돌아가서, 제어부(130)는 복수의 주파수 빈에 각각 대응하는 복수의 방향각을 군집화 하고, 군집화 결과에 기초하여 복수의 음원의 방향각 들을 획득할 수 있다(S440).

[0129] 하나의 프레임에는 오디오 신호의 복수의 주파수 빈(frequency bin)이 존재할 수 있다.

[0130] 이 경우 제어부(130)는 각각의 주파수 빈에 대하여, 앞서 설명한 처리를 통하여 방향각을 추정할 수 있다. 따라서 하나의 프레임 내 복수의 주파수 빈에 각각 대응하는 복수의 방향각이 추정될 수 있다.

[0131] 예를 들어 제어부(130)는 제1 주파수 빈에 대응하는 제1 방향각, 제2 주파수 빈에 대응하는 제2 방향각, 제n 주파수 빈에 대응하는 제n 방향각을 추정할 수 있다.

[0132] 한편 제어부(130)는 복수의 주파수 빈에 각각 대응하는 복수의 방향각을 군집화(clustering)할 수 있다. 이 경우 K 평균 클러스터링 알고리즘이 사용될 수 있다.

[0133] 그리고 제어부(130)는 군집화의 결과에 기초하여 복수의 음원의 방향각들을 획득할 수 있다.

[0134] 구체적으로 복수의 주파수 빈에 각각 대응하는 복수의 방향각은 하나 이상의 군집으로 군집화 될 수 있다.

[0135] 예를 들어 하나의 음원이 존재하며 음원의 방향각이 0도인 경우, 복수의 주파수 빈에 각각 대응하는 복수의 방향각은 0도 또는 0도 주변으로 군집화될 수 있다.

[0136] 다른 예를 들어 세개의 음원이 존재하며, 세개의 음원의 방향각이 각각 0도, -45도, +45도인 경우, 복수의 주파수 빈에 각각 대응하는 복수의 방향각은 제1 군집(0도 또는 0도 주변), 제2 군집(-45도 또는 -45도 주변), 제3 군집(+45도 또는 +45도 주변)으로 군집화 될 수 있다.

[0137] 이 경우 제어부(130)는 임계값 이하의 방향각을 제거하고, 군집에 포함되는 방향각들을 이용하여 음원의 방향각을 획득할 수 있다.

[0138] 예를 들어 총 256개의 주파수 빈 중 100개의 주파수 빈에 각각 대응하는 100개의 방향각이 제1 군집에 속하는 경우, 제어부(130)는 제1 군집에 속하는 100개의 방향각 중 적어도 하나를 이용하여 제1 대표 방향각을 획득하고, 제1 대표 방향각을 제1 음원의 방향각으로 추정할 수 있다.

[0139] 또한 총 256개의 주파수 빈 중 80개의 주파수 빈에 각각 대응하는 80개의 방향각이 제2 군집에 속하는 경우, 제어부(130)는 제2 군집에 속하는 80개의 방향각 중 적어도 하나를 이용하여 제2 대표 방향각을 획득하고, 제2 대표 방향각을 제2 음원의 방향각으로 추정할 수 있다.

[0140] 또한 총 256개의 주파수 빈 중 50개의 주파수 빈에 각각 대응하는 50개의 방향각이 제3 군집에 속하는 경우, 제어부(130)는 제3 군집에 속하는 50개의 방향각 중 적어도 하나를 이용하여 제3 대표 방향각을 획득하고, 제3 대표 방향각을 제3 음원의 방향각으로 추정할 수 있다.

[0141] 도 9는 딥 러닝 모델을 이용하여 방향각을 직접 추정하는 방식, 방향각의 정현파 함수를 추정하는 방식, 채널 간 위상 차의 정현파 함수를 추정하는 방식의 실험 결과를 도시한 도면이다.

[0142] 도 9a, 9b, 9c는 딥 러닝 모델을 이용하여 방향각을 직접 추정하는 방식을 사용했을 때의 방향각(DoA)의 분포를 도시한 도면이다. 도 9a는 단일 음원이 실제 방향각 -50도에 위치한 경우, 도 9b는 단일 음원이 실제 방향각 30도에 위치한 경우, 도 9c는 단일 음원이 실제 방향각 70도에 위치한 경우의 방향각(DoA)의 분포이다.

[0143] 또한 도 9d, 9e, 9f는 딥 러닝 모델을 이용하여 방향각의 정현파 함수(삼각 함수 벡터 특징)를 추정하는 방식을 사용했을 때의 방향각(DoA)의 분포를 도시한 도면이다. 도 9d는 단일 음원이 실제 방향각 -50도에 위치한 경우, 도 9e는 단일 음원이 실제 방향각 30도에 위치한 경우, 도 9f는 단일 음원이 실제 방향각 70도에 위치한 경우의 방향각(DoA)의 분포이다.

[0144] 또한 도 9g, 9h, 9i는 본 발명에서 제안하는 방식으로, 딥 러닝 모델을 이용하여 채널 간 위상차의 정현파 함수(삼각 함수 벡터 특징)를 추정하는 방식을 사용했을 때의 방향각(DoA)의 분포를 도시한 도면이다. 도 9g는 단일 음원이 실제 방향각 -50도에 위치한 경우, 도 9h는 단일 음원이 실제 방향각 30도에 위치한 경우, 도 9i는 단일 음원이 실제 방향각 70도에 위치한 경우의 방향각(DoA)의 분포이다.

[0145] 도 9를 참고하면, 본 발명에서 제안하는 방식에 따르는 경우, 주파수 빈들의 군집화가 월등히 잘 되어 있는 것을 알 수 있다. 따라서 본 발명에서 제안하는 방식에 따르는 경우, 음원의 방향각을 훨씬 더 정확하게 추정할 수 있다.

[0146] 도 10은 또 다른 테스트에서의 조건을 설명한 도면이다.

[0147] 훈련용 데이터는 서로 다른 제1 공간(room 1), 제2 공간(room 2), 제3 공간(room 3)에서, 다 채널 마이크의 위치(center of mic array)를 변경해 가면서, 음원과의 거리(r)를 변경해 가면서, 음원의 실제 방향 각을 -90도로

부터 +90도로 10도 간격으로 변경해 가면서, 잔향 시간(RT60)을 변경해 가면서 수집되었다.

[0148] 또한 노이즈에는, NOISEX-92 데이터베이스의 Babble, Factory 및 Volvo 노이즈가 사용되었다.

[0149] 또한 훈련용 데이터를 이용하여 학습된 딥 러닝 모델을 이용하여, 서로 다른 제4 공간(small room), 제5 공간(large room)에서, 다 채널 마이크의 위치(center of mic array)를 변경해 가면서, 음원과의 거리(r)를 변경해 가면서, 음원의 실제 방향 각을 -90도로부터 +90도로 10도 간격으로 변경해 가면서, 음원의 방향각을 추정하였다.

[0150] 도 11은 도 10의 조건에서, 노이즈를 조절해 가면서, 세가지 방향각 추정 기법을 사용하여 실험한 결과를 도시한 도면이다.

[0151] 첫번째 기법(MA)은, 선행 기술 1(N. Ma and G. J. Brown, “Speech localisation in a multitalker mixture by humans and machines,” in Proc. Interspeech, 2016, pp. 3359?3363)(N. Ma, T. May, and G. J. Brown, “Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 12, pp. 2444?2453, Dec.)에 기반한 것으로, DNN의 출력 특징(feature)의 차원(dimension)을 방향각의 분류군(class)의 수로 설정하고, DNN으로 획득하는 사후 확률(posterior probability)이 최대가 되는 클래스(class)를 해당 방향각으로 선택하는 방식이다. 예를 들어 -90~+90도 범위의 방향각을 10도 단위로 분류하면 분류군의 수는 19개가 된다.

[0152] 두번째 기법(WANG)은 선행 기술 2(Z. Q. Wang, X. Zhang, and D.-L. Wang, “Robust speaker localization guided by deep learning based time-frequency masking,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 1, pp. 178?188, Jan. 2019.)에 기반한 것으로, “Mask-weighted steered-response SNR”에 기반한 알고리즘이다.

[0153] 구체적으로 DNN regression을 통해 마스크(mask)를 추정한 후, 추정된 마스크를 이용하여 이용하여 신뢰할만한 공간정보를 가지고 있을 법한 시간-주파수 성분을 추려낸다. 그리고 나서 beamforming 기법을 접목하여, 방향각과 주파수 변화에 따른 SNR 응답(response)를 계산하고 SNR 응답(response)을 최대화 하는 최대화하는 방향각을 탐지하는 방식이다. 선행 기술 1과는 달리, DNN이 공간정보의 직접적인 추정에 활용되는 것이 아니라 가중치(weight)로 활용되는 마스크(mask) 추정에만 활용된다는 점이 특이하다.

[0154] 세번째 기법(eIPD)은 본 발명에서 제안하는, 심층 신경망 기반의 방향각 추정 방법이다.

[0155] 그리고 cIPD는, 딥 러닝 모델의 사용 없이, 채널 간 위상 차로부터 수학식 1를 이용하여 음원의 방향각을 바로 산출하는 방식을 의미할 수 있다.

[0156] 테스트는 노이즈의 종류(Babble, Factory, Volvo)를 변경해 가면서, 그리고 노이즈의 크기를 변경해가면서(SNR 5dB, SNR 10dB, SNR 15Db), 실내 공간의 크기(Large room, Small room)를 변경해가면서, 음원의 위치를 변경해 가면서 수행되었다.

[0157] 도 11a 및 도 11b를 참고하면, 본 발명에서 제안하는 세번째 기법(eIPD)은, 딥 러닝 모델을 사용하지 않은 cIPD에 비하여 월등한 성능을 나타내는 것을 알 수 있다.

[0158] 또한 본 발명에서 제안하는 세번째 기법(eIPD)은 선행 기술 2에 기반한 두번째 기법(WANG)에 비해 우수한 성능을 나타내며, 선행 기술 1에 기반한 첫번째 기법(MA)와 유사한 성능을 나타내는 것을 알 수 있다.

[0159] 도 12는 도 10의 조건에서, 잔향 시간(RT60)을 조절해 가면서, 세가지 방향각 추정 기법을 사용하여 실험한 결과를 도시한 도면이다.

[0160] 잔향 시간(RT60)을 조절해 가면서, 실내 공간의 크기(Large room, Small room)를 변경해가면서, 음원의 위치를 변경해 가면서 수행되었다.

[0161] 도 12a 및 도 12b를 참고하면, 본 발명에서 제안하는 세번째 기법(eIPD)은, 딥 러닝 모델을 사용하지 않은 cIPD, 선행 기술 1에 기반한 첫번째 기법(MA), 선행 기술 2에 기반한 두번째 기법(WANG)에 비해 훨씬 우수한 성능을 나타내는 것을 알 수 있다.

[0162] 도 13 및 도 14는, 또 다른 테스트 결과를 도시한 도면이다.

[0163] 잔향 및 노이즈가 존재하는 환경에서, 두개의 음원의 위치를 변경해 가면서, 세가지 방향각 추정 기법(THO, MA,

PROP)을 사용하여 두개의 음원의 방향각(Θ1, Θ2)을 추정하였다.

[0164] 여기서 MA는 앞서 설명한 선행 기술 1에 기반한 방향각 추정 기법이다.

[0165] 또한 THO는, 선행 기술 3(N. T. N. Tho, S. Zhao, and D. L. Jones, "RobustDoA estimation ofmultiple speech sources," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2014, pp. 2287?2291.)에 기반한 방향각 추정 기법이다.

[0166] 또한 PROP는, 본 발명에서 제안하는 방향각 추정 기법이다.

[0167] 세가지 종류의 노이즈(Babble, Factory, Volvo)를 사용하여, 그리고 두개의 잔향 시간을 사용하여, 테스트가 수행되었다.

[0168] 도 13은 실험 결과를 노이즈의 종류 별로 세분화 한 결과를 도시한 도면이고, 도 14는 실험 결과를 두개의 잔향 시간으로 세분화 한 결과를 도시한 도면이다.

[0169] 도 13을 참고하면, 노이즈의 종류(Babble, Factory, Volvo) 에 관계 없이, PROP의 평균 제곱근 오차(Root Mean Square Error, RMSE)가 가장 낮게 측정되었다.

[0170] 또한 도 14를 참고하면, 잔향의 시간(shorter RT60, longer RT60)에 관계 없이, PROP의 평균 제곱근 오차(Root Mean Square Error, RMSE)가 가장 낮게 측정되었다.

[0171] 이 것은 본 발명에서 제안하는 기법(PROP)이 가장 우수한 성능을 나타내는 것을 보여준다.

[0172] 최근, 스마트폰, 인공지능 스피커 등의 각종 모바일 기기에서 사람의 위치/방향을 탐지하는 기술은 매우 중요한 기술로써 부각되고 있다. 또한 빔포밍(beamforming) 기술은, 음원의 위치를 알고 있음을 전제로 원하는 음성이나 오디오가 위치한 방향에 집중하는 것으로, 음성향상(speech enhancement), 음성인식(speech recognition) 등의 다양한 분야에 활용될 수 있다.

[0173] 본 발명은 학습된 딥 러닝 모델을 이용하여 공간 정보, 특히 채널 간 위상 차를 직접적으로 향상시키는 방식을 취한다. 구체적으로 본 발명에서는, 왜곡된 채널 간 위상 차로부터 깨끗한 채널 간 위상 차를 추정함으로써, 기존의 기법들에 비하여 음원정위(sound source localization)의 성능을 향상시킬 수 있었다.

[0174] 또한 선행기술 1에서는, 출력 특징(output feature)을 소정의 개수의 클래스로 분류(classification)하는 방식을 취하여 분해능에 한계를 보인데 반해, 본원 발명에서는 위상 차를 삼각함수의 형태로 변환하여 회귀(regression) 모델을 생성한 후 이용하기 때문에, 분해능이 클래스의 수에 의해 제한을 받지 않는 장점이 있다.

[0175] 한편, 제어부는 일반적으로 장치의 제어를 담당하는 구성으로, 중앙처리장치, 마이크로 프로세서, 프로세서 등의 용어와 혼용될 수 있다.

[0176] 전술한 본 발명은, 프로그램이 기록된 매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 매체는, 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 읽을 수 있는 매체의 예로는, HDD(Hard Disk Drive), SSD(Solid State Disk), SDD(Silicon Disk Drive), ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 상기 컴퓨터는 단말기의 제어부(180)를 포함할 수도 있다. 따라서, 상기의 상세한 설명은 모든 면에서 제한적으로 해석되어서는 아니 되고 예시적인 것으로 고려되어야 한다. 본 발명의 범위는 첨부된 청구항의 합리적 해석에 의해 결정되어야 하고, 본 발명의 등가적 범위 내에서의 모든 변경은 본 발명의 범위에 포함된다.

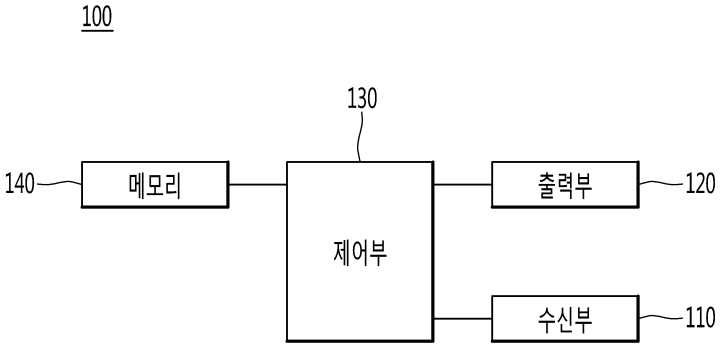
부호의 설명

[0178] 110: 수신부 120: 출력부

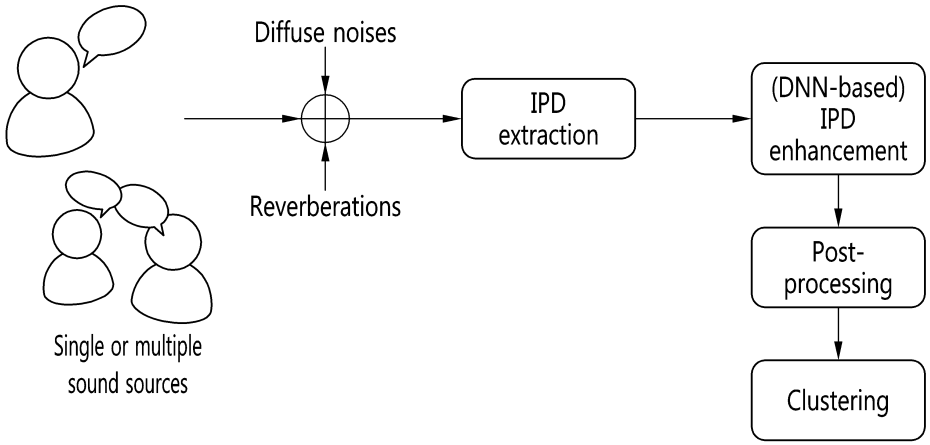
130: 제어부

도면

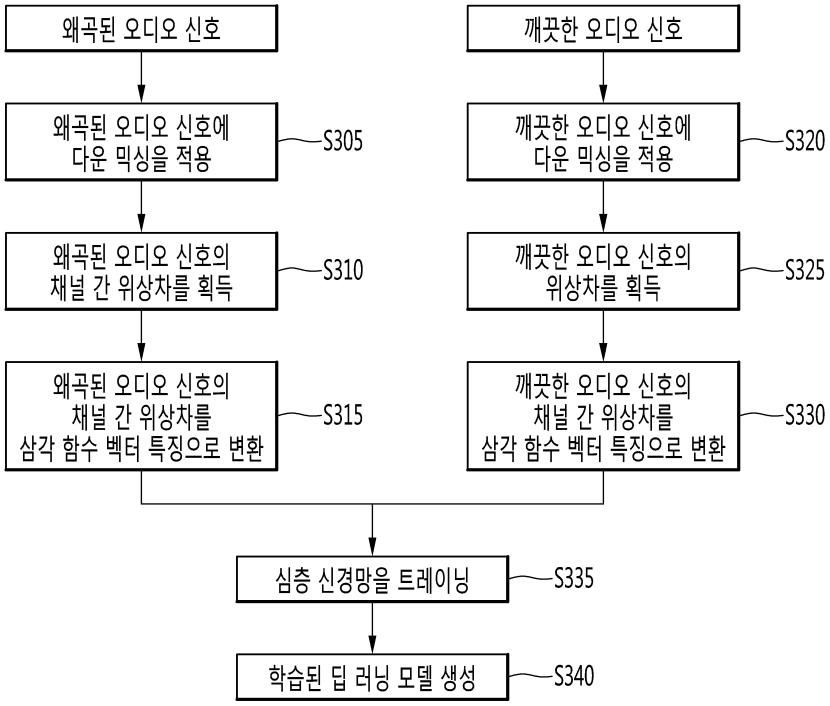
도면1


100
130
140
메모리
제어부
출력부
120
수신부
110

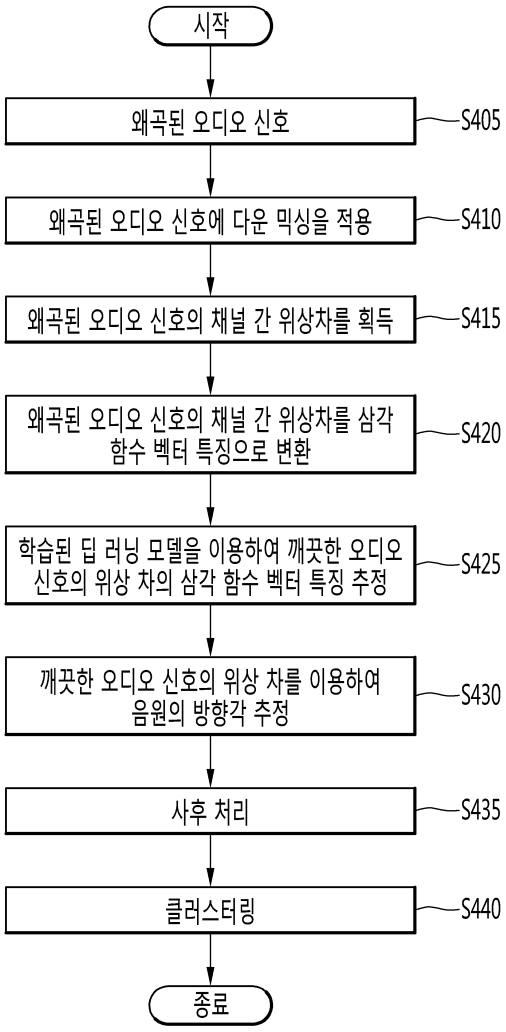
도면2


Diffuse noises
Reverberations
IPD extraction
(DNN-based) IPD enhancement
Post-processing
Clustering
Single or multiple sound sources

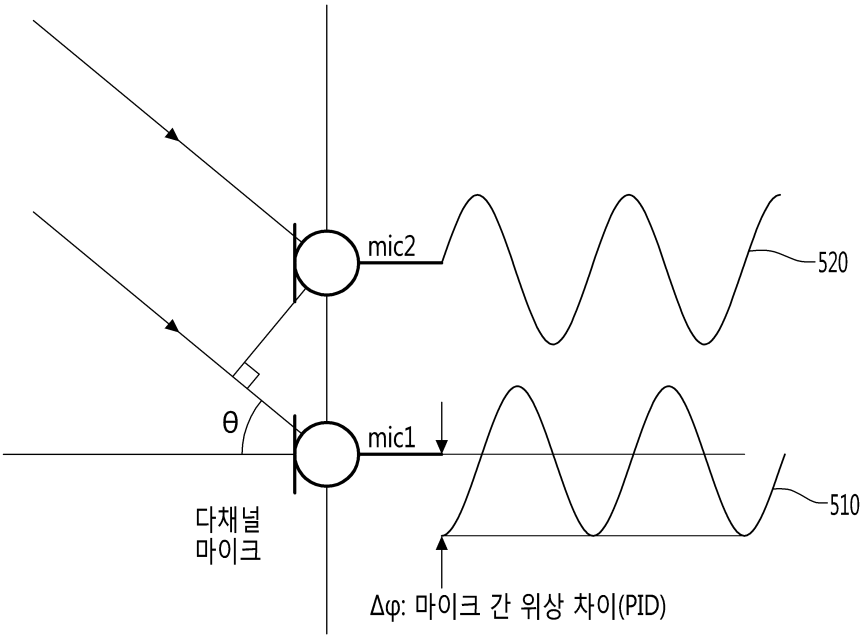
도면3


왜곡된 오디오 신호
깨끗한 오디오 신호
왜곡된 오디오 신호에 다운 믹싱을 적용
S305
깨끗한 오디오 신호에 다운 믹싱을 적용
S320
왜곡된 오디오 신호의 채널 간 위상차를 획득
S310
깨끗한 오디오 신호의 위상차를 획득
S325
왜곡된 오디오 신호의 채널 간 위상차를 삼각 함수 벡터 특징으로 변환
S315
깨끗한 오디오 신호의 채널 간 위상차를 삼각 함수 벡터 특징으로 변환
S330
심층 신경망을 트레이닝
S335
학습된 딥 러닝 모델 생성
S340

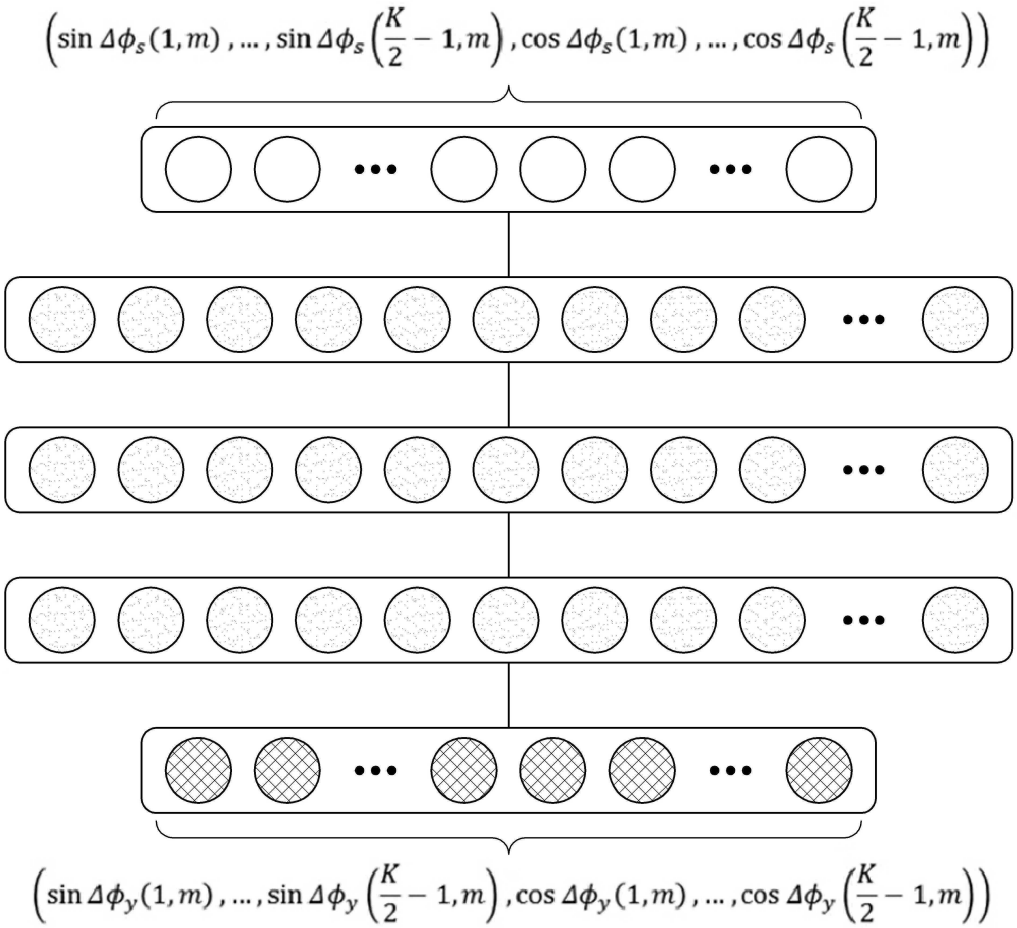
도면4


시작
왜곡된 오디오 신호
S405
왜곡된 오디오 신호에 다운 믹싱을 적용
S410
왜곡된 오디오 신호의 채널 간 위상차를 획득
S415
왜곡된 오디오 신호의 채널 간 위상차를 삼각 함수 벡터 특징으로 변환
S420
학습된 딥 러닝 모델을 이용하여 깨끗한 오디오 신호의 위상 차의 삼각 함수 벡터 특징 추정
S425
깨끗한 오디오 신호의 위상 차를 이용하여 음원의 방향각 추정
S430
사후 처리
S435
클러스터링
S440
종료

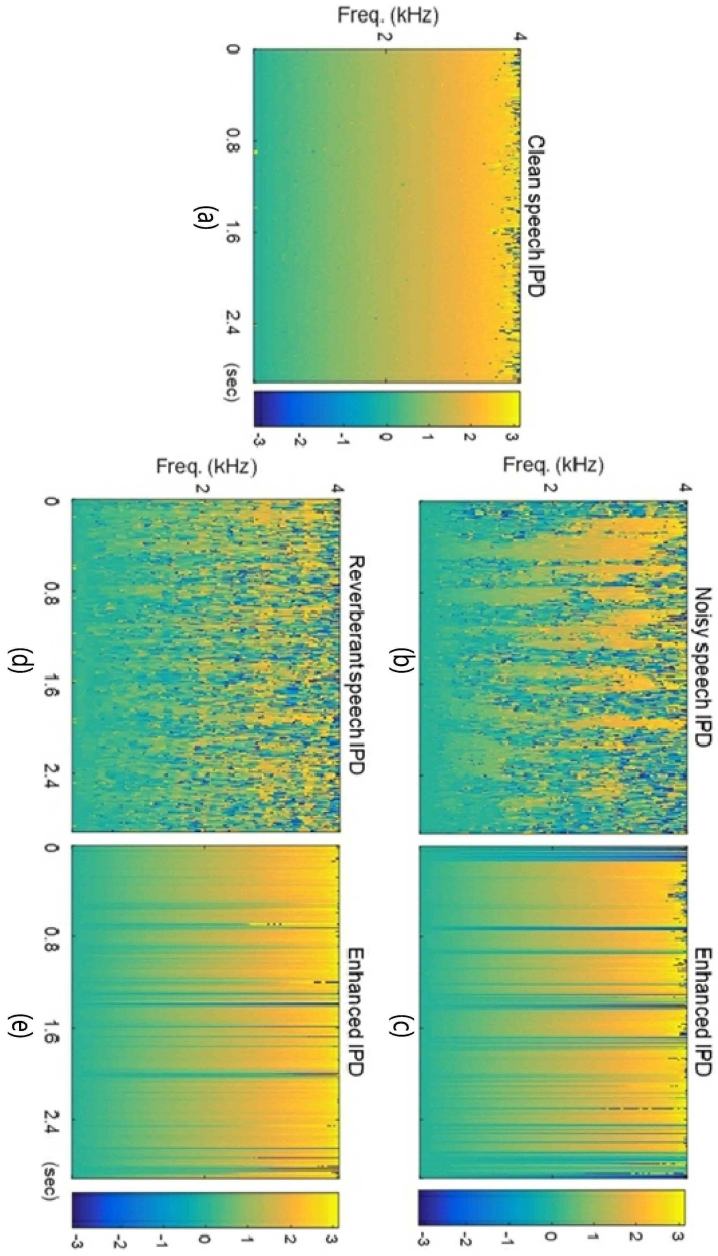
도면5


mic2
520
θ
mic1
510
다채널
마이크
Δφ: 마이크 간 위상 차이(PID)

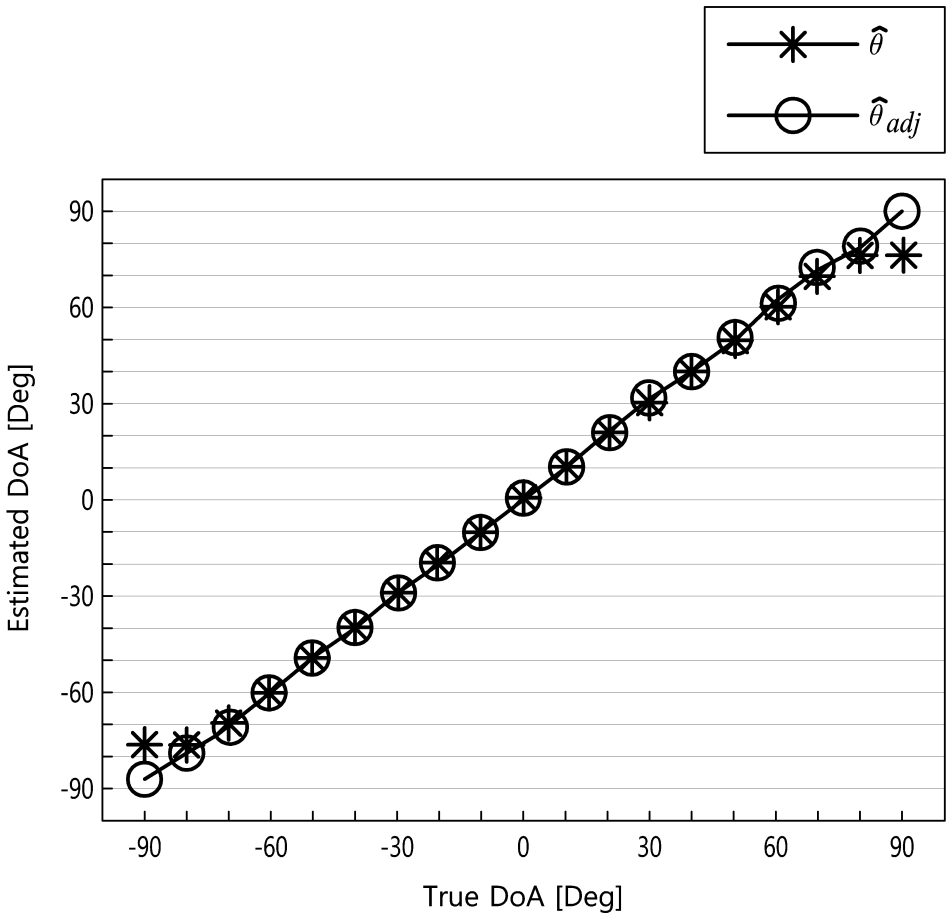
도면6


$\left(\sin\Delta\phi_s(1,m),\ldots,\sin\Delta\phi_s\left(\frac{K}{2}-1,m\right),\cos\Delta\phi_s(1,m),\ldots,\cos\Delta\phi_s\left(\frac{K}{2}-1,m\right)\right)$
$\left(\sin\Delta\phi_y(1,m),\ldots,\sin\Delta\phi_y\left(\frac{K}{2}-1,m\right),\cos\Delta\phi_y(1,m),\ldots,\cos\Delta\phi_y\left(\frac{K}{2}-1,m\right)\right)$

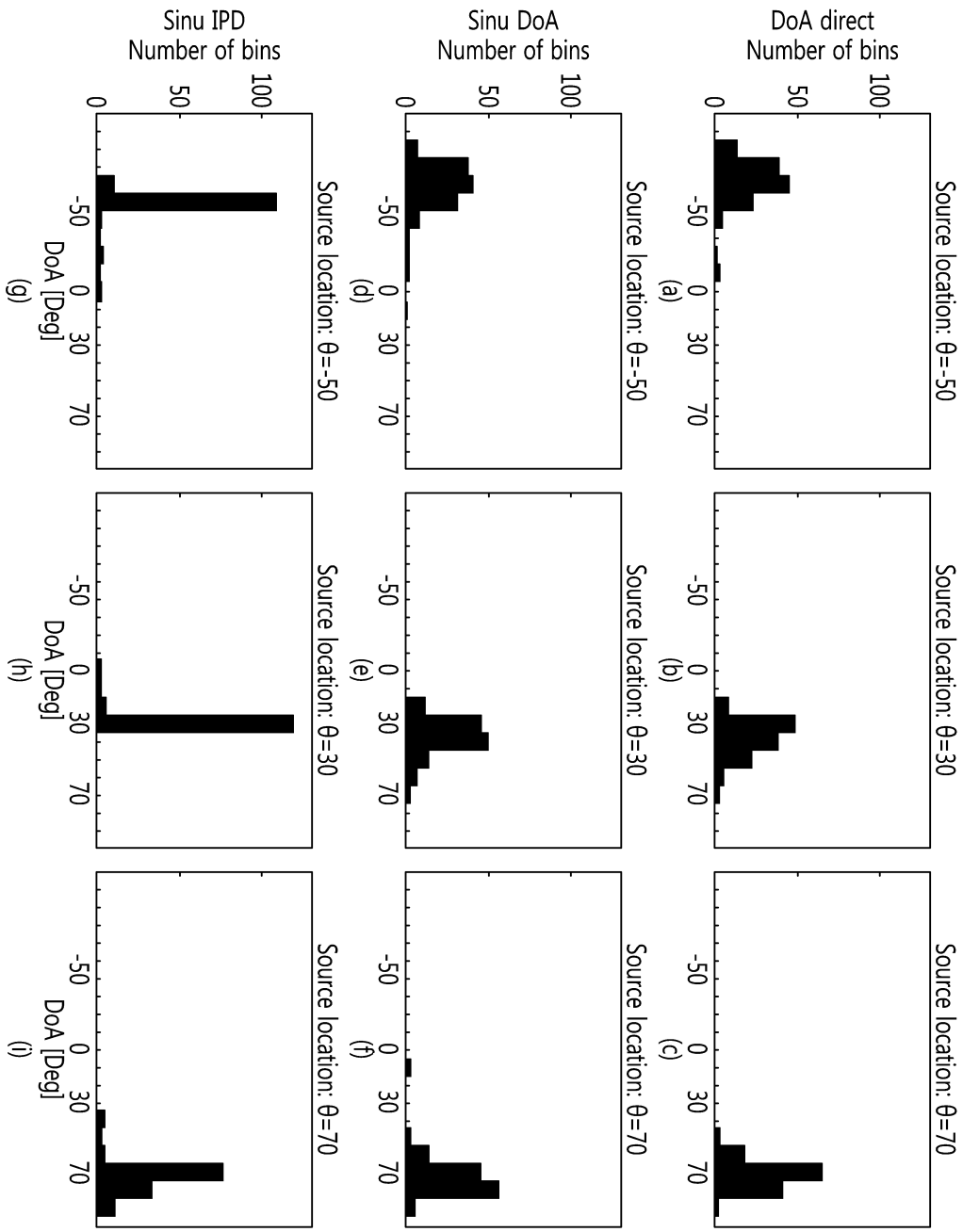
도면7


Freq. (kHz)
Clean speech IPD
(a)
Noisy speech IPD
(b)
Enhanced IPD
(c)
Reverberant speech IPD
(d)
Enhanced IPD
(e)
(sec)

도면8


$\hat{\theta}$
$\hat{\theta}_{adj}$
Estimated DoA [Deg]
True DoA [Deg]
90
60
30
0
-30
-60
-90
-90
-60
-30
0
30
60
90

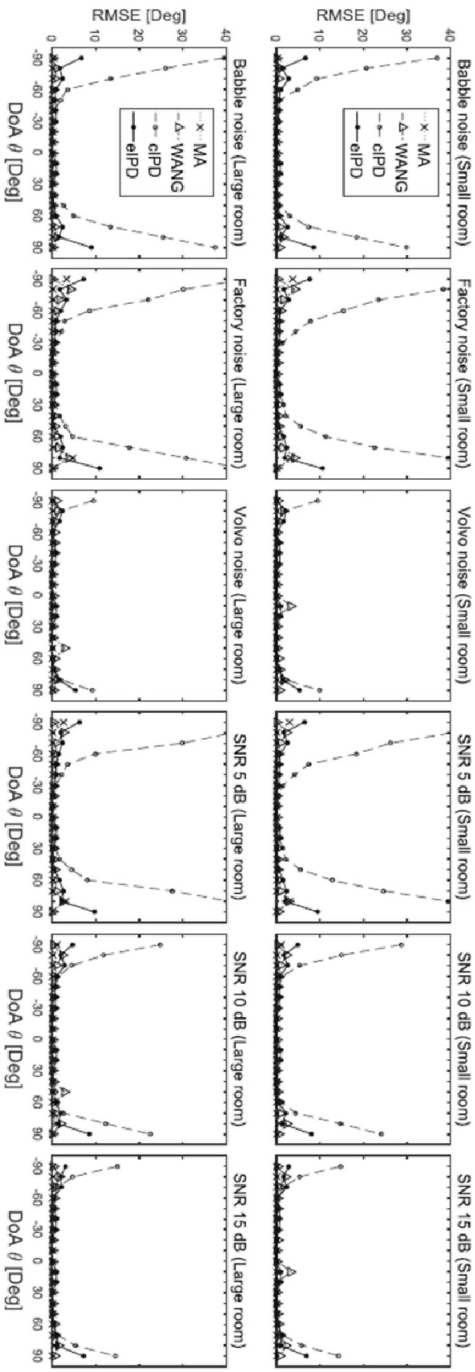
도면9


DoA direct
Number of bins
Sinu DoA
Number of bins
Sinu IPD
Number of bins
Source location: θ=-50
Source location: θ=30
Source location: θ=70
DoA [Deg]
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

도면10

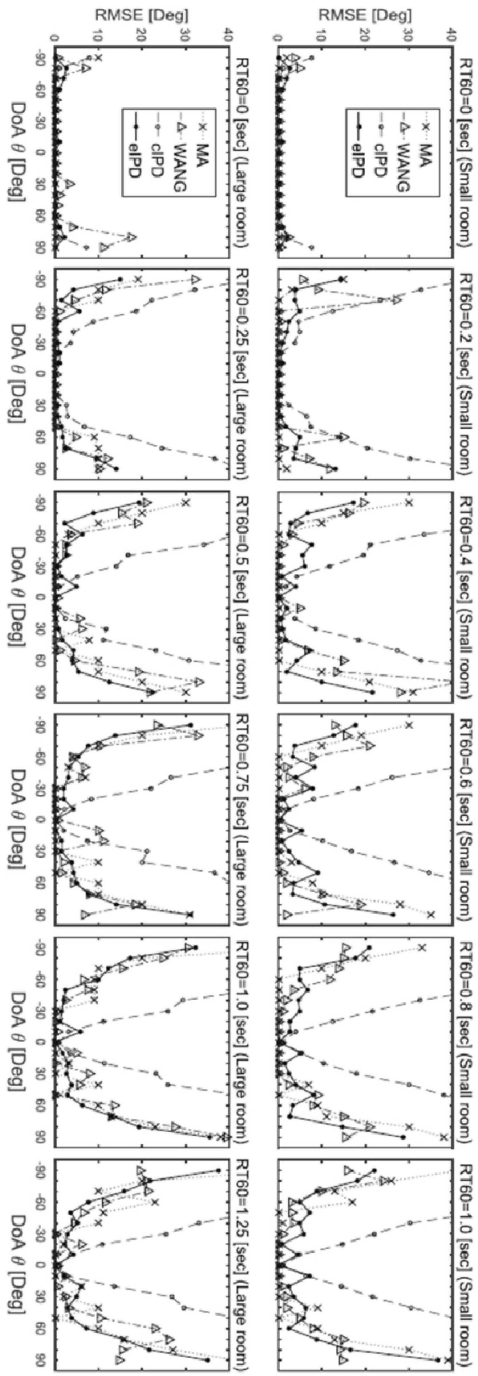
dataset	room	W(m)	L(m)	H(m)	r(m)	center of mic array (m)	RT_{60}(sec)	
							single source localization	multiple source localization
training	room 1	4.0	3.5	2.5	0.5	(2.25, 2.0, 1.5)	0 to 0.8 (interval: 0.16)	0.24, 0.64
	room 2	5.5	4.0	3.5	1.0	(2.5, 1.75, 1.6)	0 to 1.2 (interval: 0.24)	0.36, 0.72
	room 3	7.0	7.5	5.0	2.0	(3.3, 3.1, 1.7)	0 to 1.5 (interval: 0.3)	0.45, 0.9
test	small room	3.75	3.1	2.75	0.75	(1.75, 1.25, 1.55)	0 to 1.0 (interval: 0.2)	0.3, 0.7
	large room	6.0	5.0	4.0	1.5	(3.3, 1.6, 1.65)	0 to 1.25 (interval: 0.25)	0.375, 0.875

도면11



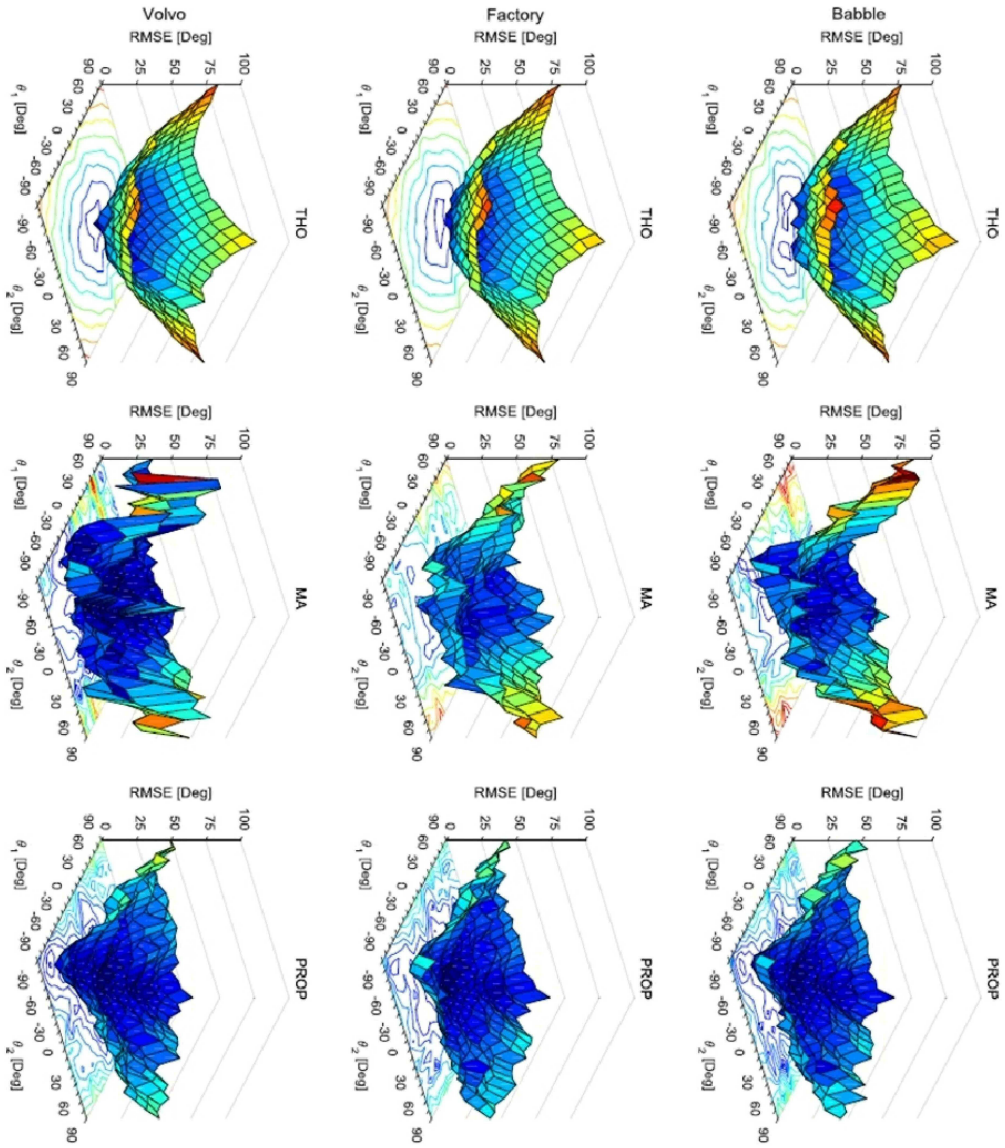
	Acc (error<15°)						Acc (error<5°)					
	Babble	Factory	Volvo	5 dB	10dB	15dB	Babble	Factory	Volvo	5 dB	10dB	15dB
cIPD	85.9	78.1	99.8	78.7	90.6	94.5	76.9	69.2	89.4	71.0	80.4	84.2
MA	100.0	100.0	100.0	100.0	100.0	100.0	99.7	96.1	100.0	96.9	99.0	99.8
WANG	99.9	100.0	99.8	100.0	99.9	99.8	99.7	96.8	98.0	98.4	98.1	98.1
PROP (eIPD)	100.0	99.4	100.0	99.5	99.9	100.0	93.7	90.8	100.0	91.4	93.2	99.8

도면12



	Acc (error<15°)						Acc (error<5°)					
	$RT60_1$	$RT60_2$	$RT60_3$	$RT60_4$	$RT60_5$	$RT60_6$	$RT60_1$	$RT60_2$	$RT60_3$	$RT60_4$	$RT60_5$	$RT60_6$
cIPD	100.0	61.8	38.9	27.9	24.7	21.6	89.5	46.1	24.5	17.1	13.9	12.1
MA	100.0	95.0	80.3	78.7	76.1	72.1	96.8	79.2	62.9	57.1	53.4	49.0
WANG	97.6	90.0	81.6	72.6	67.9	64.2	96.1	81.6	63.9	56.8	50.2	45.5
PROP (eIPD)	100.0	95.8	88.2	84.5	78.4	77.1	100.0	80.5	69.8	59.4	54.5	46.8

도면13



도면14