



(19) 대한민국특허청(KR)  
(12) 등록특허공보(B1)

(45) 공고일자 2025년04월01일  
(11) 등록번호 10-2789155  
(24) 등록일자 2025년03월26일

(51) 국제특허분류(Int. Cl.)  
G10L 21/0208 (2013.01) G10L 17/00 (2024.01)  
G10L 21/0216 (2013.01) G10L 21/0232 (2013.01)  
G10L 21/0264 (2013.01) G10L 25/78 (2013.01)  
G10L 25/90 (2013.01)  
(52) CPC특허분류  
G10L 21/0208 (2013.01)  
G10L 17/00 (2024.01)  
(21) 출원번호 10-2021-7032319  
(22) 출원일자(국제) 2019년03월10일  
심사청구일자 2022년03월04일  
(85) 번역문제출일자 2021년10월07일  
(65) 공개번호 10-2021-0137146  
(43) 공개일자 2021년11월17일  
(86) 국제출원번호 PCT/IB2019/051933  
(87) 국제공개번호 WO 2020/183219  
국제공개일자 2020년09월17일  
(56) 선행기술조사문헌  
JP2008064892 A\*  
WO2015157458 A1\*  
WO2018022222 A1\*  
Sharon Gannot et al., 'Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech', IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL.49, NO.8, AUGUST 2001.\*  
\*는 심사관에 의하여 인용된 문헌

(73) 특허권자  
카르둠 테크놀로지 엘티디.  
이스라엘, 7316000 마조르, 하게펜 30  
(72) 발명자  
스라팍, 알론  
이스라엘 마조르 7316000 하야스민 6  
체르카스스키, 다니  
이스라엘 마조르 7316000 하젠펜 30  
(74) 대리인  
특허법인에이아이피

전체 청구항 수 : 총 21 항

심사관 : 정성윤

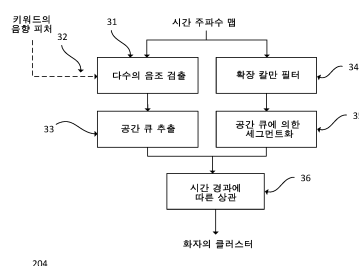
(54) 발명의 명칭 큐의 클러스터링을 사용한 음성 증강

(57) 요약

음성 증강을 위한 방법으로서, 상기 방법은 마이크로폰 어레이에 의해 주어진 시간 기간 동안 수신된 사운드 신호를 나타내는 사운드 샘플을 수신하거나 생성하는 단계; 주파수 변환된 샘플을 제공하기 위해 사운드 샘플을 주파수 변환하는 단계; 화자 관련 클러스터를 제공하기 위해 주파수 변환된 샘플을 화자들에 대하여 클러스터링하

(뒷면에 계속)

대표도 - 도3



204

는 단계로서, 클러스터링은 (i) 수신된 사운드 신호와 관련된 공간 큐 및 (ii) 화자와 관련된 음향 큐에 기초하는, 상기 클러스터링하는 단계; 화자 관련 상대 전달 함수를 제공하기 위해 화자들의 각각의 화자에 대한 상대 전달 함수를 결정하는 단계; 빔포밍된 신호를 제공하기 위해 화자 관련 상대 전달 함수에 다중 입력 다중 출력 (MIMO) 빔포밍 동작을 적용하는 단계; 및 음성 신호를 제공하기 위해 빔포밍된 신호를 역 주파수 변환하는 단계를 포함한다.

(52) CPC특허분류

*G10L 21/0232* (2013.01)

*G10L 21/0264* (2013.01)

*G10L 25/78* (2013.01)

*G10L 25/90* (2013.01)

*G10L 2021/02087* (2013.01)

*G10L 2021/02166* (2013.01)

## 명세서

### 청구범위

#### 청구항 1

컴퓨터 장치에 의해 수행될 음성 증강(speech enhancement)을 위한 방법에 있어서, 상기 방법은,

주파수 변환된 샘플을 제공하기 위해 사운드 샘플을 주파수 변환하는 단계로서, 상기 사운드 샘플은 주어진 시간 기간 동안 마이크로폰 어레이에 의해 수신된 사운드 신호를 나타내는, 상기 변환하는 단계;

상기 주파수 변환된 샘플을 복수의 화자 각각에 대응하는 복수의 화자 관련(speaker-related) 클러스터로 클러스터링하는 단계(clustering)로서, 상기 클러스터링은 상기 사운드 신호와 관련된 공간 큐(spatial cue)에 기초하고, 그리고 상기 복수의 화자와 관련된 음향 큐(acoustic cue)에 기초하는, 상기 클러스터링하는 단계;

상기 복수의 화자 관련 클러스터들을 이용하여 상기 복수의 화자 각각에 등록된 복수의 화자 관련 상대 전달 함수(relative transfer function)를 결정하는 단계 - 상기 복수의 화자 관련 상대 전달 함수를 결정하는 단계는 상기 마이크로폰 어레이에 2 개의 개별 마이크로폰에 대한 화자의 2개의 음향 전달 함수 사이의 주파수 도메인에서의 비율을 나타내는 상기 화자에 대응하는 화자 관련 상대 전달 함수를 결정하는 것을 포함함 -;

빔포밍(beamforming)된 신호를 제공하기 위해 상기 복수의 화자 관련 상대 전달 함수의 결정에 기초하여 다중 입력 다중 출력(MIMO : multiple input multiple output) 빔포밍 동작을 적용하는 단계; 및

음성 신호(speech signal)를 제공하기 위해 상기 빔포밍된 신호를 역 주파수 변환하는 단계를 포함하는, 방법.

#### 청구항 2

제1항에 있어서, 상기 음향 큐를 생성하는 단계로서,

상기 사운드 샘플에서 키워드를 검색하는 단계; 및

상기 키워드로부터 상기 음향 큐를 추출하는 단계에 의한, 상기 음향 큐를 생성하는 단계를 포함하는, 방법.

#### 청구항 3

제2항에 있어서, 상기 키워드와 관련된 공간 큐를 추출하는 단계를 포함하는, 방법

#### 청구항 4

제3항에 있어서,

상기 주파수 변환된 샘플을 상기 복수의 화자 관련 클러스터로 클러스터링하기 위한 클러스터링 시드(clustering seed)로서 상기 키워드와 관련된 상기 공간 큐를 사용하는 단계를 포함하는, 방법.

#### 청구항 5

제1항에 있어서, 상기 음향 큐는 음조(pitch) 주파수, 음조 강도, 하나 이상의 음조 주파수 고조파, 및/또는 하나 이상의 음조 주파수 고조파의 강도 중 하나 이상을 포함하는, 방법.

#### 청구항 6

제1항에 있어서, 상기 주파수 변환된 샘플을 상기 복수의 화자 관련 클러스터로 클러스터링하는 단계는 음조(pitch)의 신뢰도(reliability)를 결정하는 단계 및 상기 음조의 신뢰도가 미리 정의된 임계값 아래로 떨어질 때 상기 음조와 연관된 화자가 말이 없음(silent)이라고 결정하는 단계를 포함하는, 방법.

#### 청구항 7

제1항에 있어서, 상기 클러스터링은 상기 음향 큐 및 상기 공간 큐를 제공하기 위해 상기 주파수 변환된 샘플을 처리하는 단계; 상기 음향 큐를 사용하여 시간 경과에 따른 상기 화자의 상태를 추적하는 단계; 상기 주파수 변환된 샘플의 주파수 성분의 공간 큐를 그룹으로 세그먼트화하는 단계(segmenting); 및 활성 화자와 관련된 음향

큐를 주파수 변환된 샘플의 그룹에 할당하는 단계를 포함하는, 방법.

#### 청구항 8

제7항에 있어서, 상기 할당하는 단계는 상기 주파수 변환된 샘플의 그룹에 대해, 상기 주파수 변환된 샘플의 그룹과 관련되고, 시간 주파수 맵의 동일한 주파수 라인의 엘리먼트와 상기 시간 주파수 맵의 다른 라인에 속하는 엘리먼트 간의 교차 상관(cross-correlation)을 계산하는 단계를 포함하는, 방법.

#### 청구항 9

제7항에 있어서, 상기 추적하는 단계는 확장 칼만 필터(extended Kalman filter)를 적용하는 단계를 포함하는, 방법.

#### 청구항 10

제7항에 있어서, 상기 추적하는 단계는 다중 가설 추적(multiple hypothesis tracking)을 적용하는 단계를 포함하는, 방법.

#### 청구항 11

제7항에 있어서, 상기 추적하는 단계는 입자 필터(particle filter)를 적용하는 단계를 포함하는, 방법.

#### 청구항 12

제7항에 있어서, 상기 세그먼트화하는 단계는 단일 시간 프레임과 관련된 단일 주파수 성분을 단일 화자에게 할당하는 단계를 포함하는, 방법.

#### 청구항 13

제7항에 있어서, 상기 음향 큐를 사용하여 시간 경과에 따른 상기 화자의 상태를 추적하는 단계는 적어도 하나의 모니터링된 음향 피처(acoustic feature)를 모니터링하는 단계를 포함하고, 상기 적어도 하나의 모니터링된 음향 피처(acoustic feature)는 음성 속도, 음성 강도 및/또는 감정적 발언(emotional utterance)을 포함하는, 방법.

#### 청구항 14

제13항에 있어서, 상기 음향 큐를 사용하여 시간 경과에 따른 상기 화자의 상태를 추적하는 단계는 입력 벡터로서 상기 적어도 하나의 모니터링된 음향 피처를 확장 칼만 필터에 공급하는 단계를 포함하는, 방법.

#### 청구항 15

제1항에 있어서, 상기 주파수 변환된 샘플을 상기 복수의 화자 관련 클러스터로 클러스터링하는 단계는,

상기 주파수 변환된 샘플의 시간 주파수 맵에 따라 상기 음향 큐를 검출하기 위해 상기 주파수 변환된 샘플을 처리하는 단계;

3차원 시간 주파수 큐(time-frequency-cue) 맵에서 상기 공간 큐를 추출하기 위해 상기 주파수 변환된 샘플을 처리하는 단계; 및

상기 3차원 시간 주파수 큐 맵의 상기 음향 큐 및 상기 공간 큐에 기초하여 상기 주파수 변환된 샘플을 상기 복수의 화자 관련 클러스터에 할당하는 단계를 포함하는, 방법.

#### 청구항 16

제1항에 있어서, 상기 주파수 변환된 샘플을 클러스터링하는 단계는 상기 음향 큐를 사용하여 직접 및 간접 경로 둘 모두로부터 수신된 동일한 화자 사운드 신호에 할당하는 단계를 포함하는, 방법.

#### 청구항 17

제1항에 있어서, 상기 주파수 변환된 샘플을 상기 복수의 화자 관련 클러스터로 클러스터링하는 단계는 상기 마이크로폰 어레이의 복수의 마이크로폰 각각에 대응하는 복수의 벡터로 배열된 상기 주파수 변환된 샘플을 처리

하는 단계를 포함하고; 상기 주파수 변환된 샘플을 처리하는 단계는 상기 복수의 벡터를 가중치 평균화하여 중간 벡터를 계산하는 단계, 미리 정의된 임계값보다 낮은 값을 갖는 상기 중간 벡터의 엘리먼트를 무시함으로써 음향 큐 후보를 검색하는 단계를 포함하는, 방법.

#### 청구항 18

제17항에 있어서, 상기 미리 정의된 임계값은 노이즈의 표준 편차의 3배인, 방법.

#### 청구항 19

명령을 저장하는 비밀시적 컴퓨터 판독 가능 매체에 있어서, 상기 명령이 컴퓨터화된 시스템에 의해 실행되면 상기 컴퓨터화된 시스템이 청구항 제 1 항 내지 제 18 항 중 어느 한 항의 방법을 수행하게 하는, 비밀시적 컴퓨터 판독 가능 매체.

#### 청구항 20

마이크로폰 어레이, 메모리 유닛 및 프로세서를 포함하는 컴퓨터화된 시스템에 있어서, 상기 프로세서는 청구항 제 1 항 내지 제 18 항 중 어느 한 항의 방법을 수행하도록 구성된, 컴퓨터화된 시스템.

#### 청구항 21

청구항 제 1 항 내지 제 18 항 중 어느 한 항의 방법을 수행하도록 구성된 처리 수단을 포함하는, 장치.

#### 청구항 22

삭제

#### 청구항 23

삭제

#### 청구항 24

삭제

#### 청구항 25

삭제

#### 청구항 26

삭제

#### 청구항 27

삭제

#### 청구항 28

삭제

#### 청구항 29

삭제

#### 청구항 30

삭제

#### 청구항 31

삭제

## 청구항 32

삭제

## 청구항 33

삭제

## 청구항 34

삭제

## 청구항 35

삭제

## 청구항 36

삭제

## 청구항 37

삭제

## 발명의 설명

### 배경 기술

[0001] 음성 증강 모듈(speech enhancement module)의 성능은 원하는 음성 신호만 남기고 모든 간섭 신호를 필터링하는 능력에 달려 있다. 간섭 신호는 예를 들어, 다른 화자(speaker), 대기 상태의 노이즈, 음악, 모터 노이즈(예를 들어, 자동차 또는 비행기) 및 '칵테일 파티 노이즈(cocktail party noise)'로 알려진 대규모 군중 노이즈일 수 있다. 음성 증강 모듈의 성능은 일반적으로 음성 대 잡음비(SNR : speech-to-noise-ratio) 또는 음성 대 간섭비(SIR : speech-to-interference-ratio)를 증강시키는 능력으로 측정되며, 이는 개별적으로 노이즈 및 다른 간섭 신호의 총 파워에 대한 원하는 음성 신호의 파워의 비율(종종 dB 단위)을 반영한다.

[0002] 잔향 환경(reverberant environment)에서 음성 증강을 수행할 필요성이 증가하고 있다.

### 발명의 내용

#### 과제의 해결 수단

[0003] 음성 증강을 위한 방법이 제공될 수 있으며, 상기 방법은 : 마이크로폰 어레이에 의해 주어진 시간 기간 동안 수신된 사운드 신호를 나타내는 사운드 샘플을 수신 또는 생성하는 단계; 주파수 변환된 샘플을 제공하기 위해 사운드 샘플을 주파수 변환하는 단계; 화자 관련 클러스터를 제공하기 위해 주파수 변환된 샘플을 화자들에 대하여 클러스터링하는 단계(clustering), 상기 클러스터링은 (i) 수신된 사운드 신호와 관련된 공간 큐(spatial cue) 및 (ii) 화자와 관련된 음향 큐에 기초할 수 있는, 상기 클러스터링하는 단계; 화자 관련 상대 전달 함수를 제공하기 위해 화자들의 각각의 화자에 대한 상대 전달 함수(relative transfer function)를 결정하는 단계; 빔포밍된(beamformed) 신호를 제공하기 위해 화자 관련 상대 전달 함수에 다중 입력 다중 출력(MIMO : multiple input multiple output) 빔포밍 동작을 적용하는 단계; 및 음성 신호를 제공하기 위해 빔포밍된 신호를 역 주파수 변환하는 단계를 포함한다.

[0004] 상기 방법은 화자와 관련된 음향 큐(acoustic cue)를 생성하는 단계를 포함할 수 있다.

[0005] 상기 음향 큐를 생성하는 단계는 사운드 샘플에서 키워드를 검색하는 단계; 및

[0006] 키워드로부터 음향 큐를 추출하는 단계를 포함할 수 있다.

[0007] 상기 방법은 키워드와 관련된 공간 큐를 추출하는 단계를 포함할 수 있다.

[0008] 상기 방법은 클러스터링 시드(clustering seed)로서 키워드와 관련된 공간적 큐를 사용하는 단계를 포함할 수 있다.

- [0009] 상기 음향 큐는 음조(pitch) 주파수, 음조 강도, 하나 이상의 음조 주파수 고조파(harmonic), 및 하나 이상의 음조 주파수 고조파의 강도를 포함할 수 있다.
- [0010] 상기 방법은 신뢰도 속성을 각각의 음조에 연관시키는 단계 및 상기 음조의 신뢰도가 미리 정의된 임계값 아래로 떨어질 때 음조와 연관될 수 있는 화자가 말이 없음(silent)일 수 있다고 결정하는 단계를 포함할 수 있다.
- [0011] 상기 클러스터링은 음향 큐 및 공간 큐를 제공하기 위해 주파수 변환된 샘플을 처리하는 단계; 음향 큐를 사용하여 시간 경과에 따른 화자의 상태를 추적하는 단계; 주파수 변환된 신호의 각각의 주파수 성분의 공간 큐를 그룹으로 세그먼트화하는 단계(segmenting); 및 현재 활성(active) 화자와 관련된 음향 큐를 주파수 변환된 신호의 각각의 그룹에 할당하는 단계를 포함한다.
- [0012] 상기 할당하는 단계는 주파수 변환된 신호의 각각의 그룹에 대해, 주파수 변환 신호의 그룹과 연관될 수 있고, 시간 주파수 맵의 다른 라인에 속하는 엘리먼트와 시간 주파수 맵의 동일한 주파수 라인의 엘리먼트 사이의 교차 상관(cross-correlation)을 계산하는 단계를 포함할 수 있다.
- [0013] 상기 추적하는 단계는 확장 칼만 필터(extended Kalman filter)를 적용하는 단계를 포함할 수 있다.
- [0014] 상기 추적하는 단계는 다중 가설 추적(multiple hypothesis tracking)을 적용하는 단계를 포함할 수 있다.
- [0015] 상기 추적하는 단계는 입자 필터를 적용하는 단계를 포함할 수 있다.
- [0016] 상기 세그먼트화하는 단계는 단일 시간 프레임과 관련된 단일 주파수 성분을 단일 화자에게 할당하는 단계를 포함할 수 있다.
- [0017] 상기 방법은 음성 속도, 음성 강도 및 감정적 발언(utterance) 중 적어도 하나의 모니터링된 음향 피처를 모니터링하는 단계를 포함할 수 있다.
- [0018] 상기 방법은 적어도 하나의 모니터링된 음향 피처를 확장 칼만 필터에 공급하는 단계(feeding)를 포함할 수 있다.
- [0019] 상기 주파수 변환된 샘플은 다수의 벡터로 배열될 수 있으며, 마이크로폰 어레이의 각각의 마이크로폰당 하나의 벡터; 상기 방법은 상기 다수의 벡터를 가중치 평균화하여 중간 벡터를 계산하는 단계; 및 미리 정의된 임계값보다 낮을 수 있는 값을 가질 수 있는 중간 벡터의 엘리먼트를 무시함으로써 음향 큐 후보를 검색하는 단계를 포함한다.
- [0020] 상기 방법은 미리 정의된 임계값을 노이즈의 표준 편차의 3배로 결정하는 단계를 포함할 수 있다.
- [0021] 명령을 저장하는 비일시적 컴퓨터 판독가능 매체가 제공될 수 있고, 상기 명령이 컴퓨터화된 시스템에 의해 실행되면 상기 컴퓨터화된 시스템이 : 마이크로폰 어레이에 의해 주어진 시간 기간 동안 수신된 사운드 신호를 나타내는 사운드 샘플을 수신하거나 생성하게 하고; 주파수 변환된 샘플을 제공하기 위해 사운드 샘플을 주파수 변환하게 하고; 화자 관련 클러스터를 제공하기 위해 주파수 변환된 샘플을 화자들에 대하여 클러스터링하게 하고, 클러스터링은 (i) 수신된 사운드 신호와 관련된 공간 큐 및 (ii) 화자와 관련된 음향 큐에 기초할 수 있고; 화자 관련 상대 전달 함수를 제공하기 위해 화자들의 각각의 화자에 대한 상대 전달 함수를 결정하게 하고; 빔포밍된 신호를 제공하기 위해 화자 관련 상대 전달 함수에 다중 입력 다중 출력(MIMO) 빔포밍 동작을 적용하게 하고; 음성 신호를 제공하기 위해 빔포밍된 신호를 역 주파수 변환하게 한다.
- [0022] 비일시적 컴퓨터 판독 가능 매체는 화자와 관련된 음향 큐를 생성하기 위한 명령을 저장할 수 있다.
- [0023] 음향 큐를 생성하는 것은 사운드 샘플에서 키워드를 검색하는 것; 및
- [0024] 키워드로부터 음향 큐를 추출하는 것을 포함할 수 있다.
- [0025] 음향 큐를 생성하는 것은 사운드 샘플에서 키워드를 검색하는 것; 및
- [0026] 키워드로부터 음향 큐를 추출하는 것을 포함할 수 있다.
- [0027] 비일시적 컴퓨터 판독 가능 매체는 키워드와 관련된 공간 큐를 추출하기 위한 명령을 저장할 수 있다.
- [0028] 비일시적 컴퓨터 판독 가능 매체는 키워드와 관련된 공간 큐를 클러스터링 시드로서 사용하기 위한 명령을 저장할 수 있다.
- [0029] 음향 큐는 음조 주파수, 음조 강도, 하나 이상의 음조 주파수 고조파, 및 하나 이상의 음조 주파수 고조파의 강

도를 포함할 수 있다.

- [0030] 비밀시적 컴퓨터 판독가능 매체는 신뢰도 속성을 각각의 음조에 연관시키고, 음조의 신뢰도가 미리 정의된 임계값 아래로 떨어질 때 음조와 연관될 수 있는 화자가 말이 없음일 수 있다고 결정하기 위한 명령을 저장할 수 있다.
- [0031] 클러스터링은 음향 큐 및 공간 큐를 제공하기 위해 주파수 변환된 샘플을 처리하는 단계; 음향 큐를 사용하여 시간 경과에 따른 화자의 상태를 추적하는 단계; 주파수 변환된 신호의 각각의 주파수 성분의 공간 큐를 그룹으로 세그먼트화하는 단계; 및 현재 활성 화자와 관련된 음향 큐를 주파수 변환된 신호의 각각의 그룹에 할당하는 단계를 포함한다.
- [0032] 할당은 주파수 변환된 신호의 각각의 그룹에 대해, 주파수 변환 신호 그룹과 관련이 있고, 시간 주파수 맵의 다른 라인에 속하는 엘리먼트와 시간 주파수 맵의 동일한 주파수 라인의 엘리먼트 사이의 교차 상관을 계산하는 것을 포함할 수 있다.
- [0033] 추적은 확장 칼만 필터를 적용하는 것을 포함할 수 있다.
- [0034] 추적은 다중 가설 추적을 적용하는 것을 포함할 수 있다.
- [0035] 추적은 입자 필터를 적용하는 것을 포함할 수 있다.
- [0036] 세그먼트화는 단일 시간 프레임과 관련된 단일 주파수 성분을 단일 화자에게 할당하는 것을 포함할 수 있다.
- [0037] 비밀시적 컴퓨터 판독가능 매체는 음성 속도, 음성 강도 및 감정적 발언 중 적어도 하나의 모니터링된 음향 피처를 모니터링하기 위한 명령을 저장할 수 있다.
- [0038] 비밀시적 컴퓨터 판독가능 매체는 적어도 하나의 모니터링된 음향 피처를 확장 칼만 필터에 공급하기 위한 명령을 저장할 수 있다.
- [0039] 주파수 변환된 샘플은 다수의 벡터로 배열될 수 있으며, 마이크로폰 어레이의 각각의 마이크로폰당 하나의 벡터; 상기 비밀시적 컴퓨터 판독 가능 매체는 다수의 벡터를 가중치 평균화함으로써 중간 벡터를 계산하고; 및 미리 정의된 임계값 보다 낮을 수 있는 값을 갖는 중간 벡터의 엘리먼트를 무시함으로써 음향 큐 후보를 검색하기 위한 명령을 저장할 수 있다.
- [0040] 비밀시적 컴퓨터 판독 가능 매체는 미리 정의된 임계값을 노이즈의 표준 편차의 3배로 결정하기 위한 명령을 저장할 수 있다.
- [0041] 마이크로폰 어레이, 메모리 유닛 및 프로세서를 포함할 수 있는 컴퓨터화된 시스템이 제공될 수 있다. 프로세서는 마이크로폰 어레이에 의해 주어진 시간 기간 동안 수신된 사운드 신호를 나타내는 사운드 샘플을 수신하거나 생성하고; 주파수 변환된 샘플을 제공하기 위해 사운드 샘플을 주파수 변환하고; 화자 관련 클러스터를 제공하기 위해 주파수 변환된 샘플을 화자들에 대하여 클러스터링하고, 상기 클러스터링은 (i) 수신된 사운드 신호와 관련된 공간 큐 및 (ii) 화자와 관련된 음향 큐에 기초할 수 있고; 화자 관련 상대 전달 함수를 제공하기 위해 화자들의 각각의 화자에 대한 상대 전달 함수를 결정하고; 빔포밍된 신호를 제공하기 위해 화자 관련 상대 전달 함수에 다중 입력 다중 출력(MIMO) 빔포밍 동작을 적용하고; 음성 신호를 제공하기 위해 빔포밍된 신호를 역 주파수 변환하도록 구성될 수 있고; 및 상기 메모리 유닛은 상기 사운드 샘플 및 상기 음성 신호 중 적어도 하나를 저장하도록 구성될 수 있다.
- [0042] 상기 컴퓨터화된 시스템은 마이크로폰 어레이를 포함하지 않을 수 있지만 마이크로폰 어레이에 의해 주어진 시간 기간 동안 수신된 사운드 신호를 나타내는 마이크로폰 어레이로부터 신호를 수신할 수 있다.
- [0043] 상기 프로세서는 화자와 관련된 음향 큐를 생성하도록 구성될 수 있다.
- [0044] 상기 음향 큐를 생성하는 것은 사운드 샘플에서 키워드를 검색하는 것; 및
- [0045] 키워드로부터 음향 큐를 추출하는 것을 포함할 수 있다.
- [0046] 상기 프로세서는 키워드와 관련된 공간 큐를 추출하도록 구성될 수 있다.
- [0047] 상기 프로세서는 클러스터링 시드로서 키워드와 관련된 공간 큐를 사용하도록 구성될 수 있다.
- [0048] 상기 음향 큐는 음조 주파수, 음조 강도, 하나 이상의 음조 주파수 고조파, 및 하나 이상의 음조 주파수 고조파의 강도를 포함할 수 있다.

- [0049] 상기 프로세서는 신뢰도 속성을 각각의 음조에 연관시키고, 음조의 신뢰도가 미리 정의된 임계값 아래로 떨어질 때 음조와 연관될 수 있는 화자가 말이 없음일 수 있다고 결정하도록 구성될 수 있다.
- [0050] 상기 프로세서는 음향 큐 및 공간 큐를 제공하기 위해 주파수 변환된 샘플을 처리함으로써 클러스터링하고; 음향 큐를 사용하여 시간 경과에 따른 화자의 상태를 추적하고; 주파수 변환된 신호의 각각의 주파수 성분의 공간 큐를 그룹으로 세그먼트화하고; 현재 활성 화자와 관련된 음향 큐를 주파수 변환된 신호의 각각의 그룹에 할당하도록 구성될 수 있다.
- [0051] 상기 프로세서는 주파수 변환된 신호의 각각의 그룹에 대해, 주파수 변환된 신호 그룹과 관련될 수 있고, 시간 주파수 맵의 다른 라인에 속하는 엘리먼트와 시간 주파수 맵의 동일한 주파수 라인의 엘리먼트 간의 교차 상관을 계산함으로써 할당하도록 구성될 수 있다.
- [0052] 상기 프로세서는 확장된 칼만 필터를 적용함으로써 추적하도록 구성될 수 있다.
- [0053] 상기 프로세서는 다중 가설 추적을 적용함으로써 추적하도록 구성될 수 있다.
- [0054] 상기 프로세서는 입자 필터를 적용함으로써 추적하도록 구성될 수 있다.
- [0055] 상기 프로세서는 단일 시간 프레임과 관련된 단일 주파수 성분을 단일 화자에게 할당함으로써 세그먼트화하도록 구성될 수 있다.
- [0056] 상기 프로세서는 음성 속도, 음성 강도 및 감정적 발언 중 적어도 하나의 모니터링된 음향 피치를 모니터링하도록 구성될 수 있다.
- [0057] 상기 프로세서는 적어도 하나의 모니터링된 음향 피치를 확장 칼만 필터에 공급하도록 구성될 수 있다.
- [0058] 상기 주파수 변환된 샘플은 다수의 벡터로 배열될 수 있으며, 마이크로폰 어레이의 각각의 마이크로폰당 하나의 벡터; 상기 프로세서는 다수의 벡터를 가중 평균함으로써 중간 벡터를 계산하고; 미리 정의된 임계값 보다 낮을 수 있는 값을 가질 수 있는 중간 벡터의 엘리먼트를 무시함으로써 음향 큐 후보를 검색하도록 구성될 수 있다.
- [0059] 상기 프로세서는 미리 정의된 임계값을 노이즈의 표준 편차의 3배로 결정하도록 구성될 수 있다.

### 도면의 간단한 설명

- [0060] 본 발명을 이해하고 그것이 실제로 어떻게 수행될 수 있는지 알기 위해, 첨부 도면을 참조하여 비제한적인 예로서 바람직한 실시예가 이제 설명될 것이다.
- 도 1은 다중 경로를 예시한다.
- 도 2는 방법의 예를 예시한다.
- 도 3은 도 2의 방법의 클러스터링 단계의 예를 예시한다.
- 도 4는 시간 주파수 맵(time-frequency map)에 대한 음조(pitch) 검출의 예를 예시한다.
- 도 5는 시간 주파수 큐 맵의 예를 예시한다.
- 도 6은 오프라인 트레이닝에서 음성 인식 체인의 예를 예시한다.
- 도 7은 실시간 트레이닝에서 음성 인식 체인의 예를 예시한다.
- 도 8은 트레이닝 메커니즘의 예를 예시한다.
- 도 9는 방법의 예를 예시한다.

### 발명을 실시하기 위한 구체적인 내용

- [0061] 시스템에 대한 임의의 참조는 시스템에 의해 실행되는 방법 및/또는 시스템에 의해 일단 실행되면 시스템이 방법을 실행하도록 하는 명령을 저장하는 비일시적 컴퓨터 판독 가능 매체에 준용하여 적용되어야 한다.
- [0062] 방법에 대한 임의의 참조는 방법을 실행하도록 구성된 시스템 및/또는 시스템에 의해 일단 실행되면 시스템이 방법을 실행하도록 하는 명령을 저장하는 비일시적 컴퓨터 판독 가능 매체에 준용하여 적용되어야 한다.
- [0063] 비일시적 컴퓨터 판독가능 매체에 대한 임의의 참조는 시스템에 의해 실행되는 방법 및/또는 비일시적 컴퓨터

판독가능 매체에 저장된 명령을 실행하도록 구성된 시스템에 준용하여 적용되어야 한다.

[0064]

"및/또는"이라는 용어는 추가적이거나 또는 대안적이다.

[0065]

"시스템"이라는 용어는 컴퓨터화된 시스템을 의미한다.

[0066]

음성 증강 방법은 신호가 노이즈 및 다른 화자에 의해 간섭을 받는 경우 원하는 소스(화자)로부터 음성 신호를 추출하는 데 초점이 맞춰져 있다. 자유로운 환경에서는, 방향성 빔포밍 형태의 공간 필터링이 효과적이다. 그러나 잔향 환경에서, 각각의 소스의 음성은 반드시 연속적이지 않은 여러 방향에 걸쳐 손상되기 때문에 통상의 빔형성기(beamformer)의 장점이 저하된다. 이 문제를 해결하기 위해 전달 함수(TF : transfer-function) 기반 빔형성기를 사용하거나 TF 자체로 상대 전달 함수(RTF : relative transfer function)를 사용하는 것이 유망한 방향이다. 그러나 다수의 화자 환경에서 음성 신호가 동시에 캡처될 때 각각의 화자의 RTF를 추정하는 능력은 아직 어려운 문제이다. 동시에 화자들을 클러스터링하여 잔향 환경에서 화자들의 RTF 추정을 가능하게 하기 위해 음향 큐 및 공간 큐를 추적하는 것을 포함하는 솔루션이 제공된다.

[0067]

특히 다수의 화자 잔향 환경에서 원래 화자에 각각의 주파수 성분을 할당하는 화자의 클러스터링 알고리즘(clustering algorithm)이 제공된다. 이것은 RTF 추정기가 다수의 화자 잔향 환경에서 제대로 작동하는 데 필요한 조건을 제공한다. 그런 다음, RTF 행렬의 추정을 사용하여 TF-LCMV(transfer function based linear constrained minimum variance) 빔형성기의 가중치 벡터를 계산하고(아래의 방정식 (10) 참조) TF-LCMV가 작동하는데 필요한 조건을 충족한다. 각각의 사람의 화자에게는 상이한 음조가 부여되어 음조가 화자에 대한 전단사 표시자(bijective indicator)라고 가정된다. 다수의 음조 검출은 특히 시끄럽고 잔향이 많은 다수의 화자 환경에서 어려운 작업으로 알려져 있다. 이 문제를 해결하기 위해 W-DO(W-Disjoint Orthogonality) 가정이 사용되었으며, 신호 강도, 방위각 및 양각(elevation angle)과 같은 공간 큐 세트가 추가 피쳐로 사용된다. 음향 큐 - 음조 값 - 는 확장된 칼만 필터(EKF : extended Kalman filter)를 사용하여 시간 경과에 따라 추적되어 일시적인 비활성 화자와 음조의 변화를 극복하고, 공간 큐는 마지막 L 주파수 성분을 세그먼트화하고 각각의 주파수 성분을 다른 소스에 할당하는 데 사용된다. EKF와 세그먼트화의 결과는 교차 상관을 통해 조합되어 특정 음조의 특정 화자에 대한 주파수 성분의 클러스터링을 가능하게 한다.

[0068]

도 1은 음성 신호의 주파수 성분이 잔향 환경에서 사람 화자(11)로부터 마이크로폰 어레이(12)로 이동하는 경로를 설명한다. 환경(14)의 벽(13) 및 기타 엘리먼트는 벽의 재료 및 질감에 의존하는 감쇠 및 반사 각도로 충돌 신호를 반사한다. 사람 음성의 다른 주파수 성분은 다른 경로를 취할 수 있다. 이들은 사람 화자(11)와 마이크로폰 어레이(12) 사이의 최단 경로에 있는 직접 경로(15)이거나 간접 경로(16, 17)일 수 있다. 주파수 성분은 하나 이상의 경로를 따라 이동할 수 있다.

[0069]

도 2는 알고리즘을 설명한다. 신호는  $M \geq 2$  마이크로폰을 포함하는 마이크로폰 어레이(201)에 의해 획득되며, 여기서  $M=7$  마이크로폰은 하나의 예이다. 마이크로폰은 직선, 원 또는 구에 균등한 간격을 두거나 임의의 형상을 형성하는 고르지 않은 간격과 같은 다양한 집합체(constellation)에 배치될 수 있다. 각각의 마이크로폰으로부터의 신호는 샘플링되고, 디지털화되고, M 프레임에 저장되며, 각각은 T개의 연속적인 샘플(202)을 포함한다. 프레임 T의 크기는 STFT(short-time Fourier transform)가 정확하도록 충분히 크게 선택될 수 있지만, 그러나 신호가 등가 시간 동안 정적(stationary)일 정도로 충분히 짧다. T의 전형적인 값은 16kHz의 샘플링 속도에 대해 4,096개의 샘플이고, 즉, 프레임은 1/4초에 해당한다. 종종 연속 프레임은 시간 경과에 따라 신호의 피쳐(feature)를 추적하기 위해 서로 중첩된다. 전형적인 중첩은 75%이고, 즉, 1,024개 샘플마다 새 프레임이 시작된다. 예를 들어, T는 0.1Sec - 2Sec 사이의 범위일 수 있고, 따라서 16kHz 샘플링 속도에 대해 샘플링된 1024-32768을 제공할 수 있다. 샘플은 T 기간 동안 마이크로폰의 어레이에 의해 수신된 사운드 신호를 나타내는 사운드 샘플이라고도 한다.

[0070]

각각의 프레임은 (203)에서 푸리에 변환 또는 푸리에 변환의 변형 예컨대, STFT(short time Fourier transform), CQT(constant-Q transform), LFT(logarithmic Fourier transform), 필터 बैं크 등을 적용하여 주파수 도메인으로 변환된다. 프레임링 효과(framing effect)를 제어하기 위해 윈도우링(windowing) 및 제로 패딩(zero-padding)과 같은 여러 기술이 적용될 수 있다. (203)의 결과는 길이가 K인 M개의 복소수 값 벡터이다. 예를 들어, 어레이에 7개의 마이크로폰이 포함된 경우, 프레임 시간 인덱스  $\ell$ 에 의해 등록된 7개의 벡터가 준비된다. K는 주파수 빈(frequency bin)의 수이며 주파수 변환에 의해 결정된다. 예를 들어, 일반 STFT를 사용할 때,  $K = T$ 는 버퍼의 길이이다. 단계 (203)의 출력은 주파수 변환된 신호로 지칭될 수 있다.

[0071]

음성 신호는 (204)에서 상이한 화자에 클러스터링된다. 클러스터는 화자 관련 클러스터(speaker related

cluster)로 지칭될 수 있다. 방향에만 기초하여 화자를 클러스터링하는 종래 기술 작업과 달리, (204)는 잔향 환경에서 다수의 화자를 다루어서 직접 경로와 간접 경로로 인해 상이한 방향의 신호가 동일한 화자에 할당될 수 있다. 제안된 솔루션은 마이크로폰 중 하나에서 신호의 방향(방위각 및 앙각) 및 강도와 같은 공간 큐 세트 위에 음향 큐 세트 예를 들어, 음조 주파수 및 강도, 고조파 주파수 및 강도를 사용할 것을 제안한다. 음조 및 하나 이상의 공간 큐는 칼만 필터 및 그 변형, 다중 가설 추적(MHT : multiple hypothesis tracking) 또는 입자 필터(particle filter)와 같은 추적 알고리즘의 상태 벡터로 사용되고, 이들은 이 상태 벡터(state vector)를 추적하고 각각의 트랙을 다른 화자에 할당하는 데 사용된다.

[0072] 이러한 모든 추적 알고리즘은 시간에 따라 상태 벡터의 역학을 설명하는 모델을 사용하므로 상태 벡터의 측정치가 노이즈에 의해 누락되거나 손상될 때 추적 알고리즘은 동적 모델을 사용하여 이를 보상하고 동시에 모델 파라미터를 업데이트한다. 이 스테이지의 출력은 주어진 시간  $\ell$ 에 각각의 화자에 각각의 주파수 성분을 할당하는 벡터이다. (204)는 도 3에서 더 자세히 설명된다.

[0073] (205)에서 주파수 도메인의 데이터에 RTF 추정기가 적용된다. 이 스테이지의 결과는 각각이 관련 화자에게 등록된 일련의 RTF이다. 등록 프로세스는 클러스터링 화자(204)로부터 클러스터링 어레이를 사용하여 수행된다. RTF 세트는 화자 관련 상대 전달 함수라고도 한다.

[0074] MIMO 빔 형성기(beamformer)(206)는 공간 필터링에 의해 요구되는 음성 신호의 에너지에 대해 노이즈 및 간섭 신호의 에너지를 감소시킨다. 단계(206)의 출력은 빔포밍된 신호로 지칭될 수 있다. 그런 다음 빔포밍된 신호는 역 주파수 변환(207)으로 포워드되어 샘플의 스트림(stream)의 형태로 연속 음성 신호를 생성하고, 이는 결국 음성 인식, 통신 시스템 및 레코딩 디바이스(208)와 같은 다른 엘리먼트로 전송된다.

[0075] 본 발명의 바람직한 실시예에서, 키워드 스포팅(keyword spotting)(209)은 클러스터링 블록(204)의 성능을 개선하기 위해 사용될 수 있다. (202)로부터의 프레임은 미리 정의된 키워드(예를 들어, "헬로우 알렉사(hello Alexa)", 또는 "오케이 구글")를 검색한다. 프레임 스트림에서 키워드가 스포팅되면, 음조 주파수 및 강도, 고조파 주파수 및 강도와 같은 화자의 음향 큐가 추출된다. 또한, 각각의 주파수 성분이 마이크로폰 어레이(201)에 도달한 경로의 피치가 추출된다. 이들 피치는 클러스터링 화자(204)에 의해 원하는 화자의 클러스터에 대한 시드(seed)로서 사용된다. 시드는 클러스터의 초기 파라미터에 대한 초기 추측이다. 예를 들어, 클러스터의 중심, 반경 및 중심 기반 클러스터링 알고리즘에 대한 통계 예컨대, K-평균, PSO 및 2KPM. 또 다른 예는 서브 공간 기반 클러스터링을 위한 서브 공간의 베이스(base)이다.

[0076] 도 3은 화자의 클러스터링 알고리즘을 설명한다. 각각의 화자에는 상이한 음향 큐 세트, 예를 들어, 음조 주파수 및 강도, 고조파 주파수 및 강도가 부여되어 음향 큐 세트가 화자에 대한 전단사 표시자라고 가정된다. 음향 큐 검출은 특히 시끄럽고 잔향이 많은 다수의 화자 환경에서 어려운 작업으로 알려져 있다. 이 문제를 해결하기 위해, 신호 강도, 방위각 및 앙각의 형태에 공간 큐가 사용된다. 일시적인 비활성 화자와 음향 큐의 변화를 극복하기 위해 입자 필터 및 확장 칼만 필터(EKF)와 같은 필터를 사용하여 시간 경과에 따라 음향 큐가 추적되고, 공간 큐는 여러 소스 간의 주파수 성분을 세그먼트화하는 데 사용된다. EKF와 세그먼트화의 결과는 교차 상관을 통해 조합되어 특정 음조의 특정 화자에 대한 주파수 성분의 클러스터링을 가능하게 한다.

[0077] (31)에서 음조 주파수 형태의 잠재적인 음향 큐가 바람직한 실시예의 예로서 검출된다. 먼저 (203)에서 계산된 각각의 마이크로폰으로부터의 버퍼의 주파수 변환을 사용하여 시간 주파수 맵(time-frequency map)을 준비한다. 다음으로,  $M \times K$ -길이 복소수 값 벡터 각각의 절대값은 가중치 평균화되고, 일부 가중치 인자가 마이크로폰에서 아티팩트(artifact)를 줄이기 위해 결정될 수 있다. 결과는 단일  $K$ -길이 실수 벡터(real vector)이다. 이 벡터에서, 주어진 임계값  $\mu$ 보다 높은 값이 추출되고, 나머지 엘리먼트는 폐기된다. 임계값  $\mu$ 는 종종 노이즈의 표준 편차의 3배로 적응적으로 선택되지만, 시스템의 전기적 파라미터, 특히 샘플링된 신호의 유효 비트 수에 따라 달라지는 상수 값 이상이어야 한다.  $[k_{min}, k_{max}]$  범위의 주파수 인덱스를 갖는 값이 음조 주파수의 후보로 정의된다. 일반적인 성인 남성의 기본 주파수는 85 내지 1800Hz이고 일반적인 성인 여성의 기본 주파수는 165 내지 2550Hz이기 때문에 변수  $k_{min}$  및  $k_{max}$ 는 전형적으로 각각 85Hz 및 2550Hz이다. 그런 다음 각각의 음조 후보는 더 높은 고조파를 검색하여 검증된다. 2차 및 3차 고조파의 존재는 후보 음조가 신뢰도  $R$ (즉,  $R=10$ )을 갖는 합법적인 음조로 검출되기 위한 전제 조건일 수 있다. 더 높은 고조파(예를 들어, 4차 및 5차)가 존재하는 경우, 음조의 신뢰도가 증가될 수 있다 - 예를 들어, 각각의 고조파에 대해 2배. 일 예는 도 4에서 찾을 수 있다. 본 발명의 바람직한 실시예에서, 원하는 화자(32)의 음조는 원하는 화자에 의해 발언된 키워드를 사용하여 (210)에 의해 제공된다. 제공된 음조(32)는 가능한 가장 높은 신뢰도(예를 들어,  $R=1000$ )로 리스트에 추가된다.

[0078] (33)에서, 확장 칼만 필터(EKF)는 (31)로부터 음조에 적용된다. 확장 칼만 필터에 대한 위키피디아 항목([www.wikipedia.org/wiki/Extended\\_Kalman\\_filter](http://www.wikipedia.org/wiki/Extended_Kalman_filter))에 의해 언급된 바와 같이, 칼만 필터는 상태 전이 방정식(state transition equation)과 관찰 모델(observation model)을 갖는다. 이산 계산(discrete calculation)에 대한 상태 전이 방정식은 다음과 같다:

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_k) + \mathbf{w}_k \quad (1)$$

[0079] 그리고 이산 계산에 대한 관찰 모델은 다음과 같다:

$$\mathbf{z}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{v}_k \quad (2)$$

[0081] 여기서  $\mathbf{x}_k$ 는 시스템의 상태를 (부분적으로) 설명하는 파라미터를 포함하는 상태 벡터이고,  $\mathbf{u}_k$ 는 시스템 상태에 대한 정보를 제공하는 외부 입력의 벡터이고,  $\mathbf{w}_k$  및  $\mathbf{v}_k$ 는 프로세스 노이즈 및 관찰 노이즈이다. 확장 칼만 필터의 시간 업데이트는 예측 방정식(prediction equation)으로 다음 상태를 예측할 수 있고 검출된 음조는 다음 유형의 방정식을 사용하여 실제 측정과 예측된 측정을 비교하여 변수를 업데이트할 수 있다 :

$$\mathbf{y}_k = \mathbf{z}_k - \mathbf{h}(\mathbf{x}_{k|k+1}) \quad (3)$$

[0083] 여기서,  $\mathbf{z}_k$ 는 검출된 음조이고,  $\mathbf{y}_k$ 는 측정치와 예측된 음조 사이의 오류다.

[0084] (33)에서, 각각의 궤적(trajjectory)은 검출된 음조로부터 시작하고, 감정 때문에 더 높거나 낮을 수 있는 음조의 시간적 거동을 반영하는 모델  $f(\mathbf{x}_k, \mathbf{u}_k)$ 가 뒤따를 수 있다. 모델의 입력은 과거 상태 벡터  $\mathbf{x}_k$ (하나 이상의 상태 벡터) 및 음성의 속도, 음성의 강도 및 감정적 발언과 같은 음조의 역학에 영향을 미치는 임의의 외부 입력  $\mathbf{u}_k$ 일 수 있다. 상태 벡터  $\mathbf{x}$ 의 엘리먼트는 음조를 정량적으로 설명할 수 있다. 예를 들어, 음조의 상태 벡터는 특히 음조 주파수, 1차 고조파의 강도, 더 높은 고조파의 주파수 및 강도를 포함할 수 있다. 벡터 함수  $f(\mathbf{x}_k, \mathbf{u}_k)$ 는 현재 시간보다 앞서 어떤 주어진 시간  $k+1$ 에서 상태 벡터  $\mathbf{x}$ 를 예측하는 데 사용될 수 있다. EKF에서 동적 모델의 예시적인 구현은 Jerry M. Mendel의 책 "Lessons in Digital Estimation Theory"에 설명된 시간 업데이트 방정식(예측 방정식이라고도 함)을 포함할 수 있으며, 이는 본 출원에 참조로 통합된다.

[0085] 예를 들어, 3-튜플(tuple) 상태 벡터를 고려하면:

$$\mathbf{x}_k = \begin{bmatrix} f_k & a_k & b_k \end{bmatrix}^T \in \mathbb{R}^3 \quad (4)$$

[0087] 여기서,  $f_k$ 는 시간  $k$ 에서 음조(1차 고조파)의 주파수이고,  $a_k$ 은 시간  $k$ 에서 음조(1차 고조파)의 강도이고,  $b_k$ 는 시간  $k$ 에서 2차 고조파의 강도이다.

[0088] 음조에 대한 예시적인 상태 벡터 모델은 다음과 같을 수 있다 :

$$\mathbf{x}_k = \mathbf{x}_{k-1} \in \mathbb{R}^4 \quad (5)$$

[0089] 이것은 항상 일정한 음조를 가정하는 모델을 설명한다. 본 발명의 바람직한 실시예에서, 당업계에 공지된 음성 인식 알고리즘을 사용하는 음성의 속도, 음성의 강도 및 감정적 발언은 EKF의 시간 업데이트 스테이지를 개선하는 외부 입력  $\mathbf{u}_k$ 을 제공하여 지속적으로 모니터링된다. 감정적 발언 방법은 당업계에 알려져 있다. 예를 들어, Palo et al.의 "New Features for Emotional Speech Recognition" 참조.

[0090] 각각의 트랙에는 시간 업데이트만을 사용하여 트랙이 진화하는 시간에 반비례하는 신뢰도 필드(reliability field)가 부여된다. 트랙의 신뢰도가 검출되지 않은 음조의 10초를 나타내는 일부 신뢰도 임계값  $\rho$  아래로 내려 가면 트랙은 데드(dead)로 정의되며, 이는 해당 화자가 활성 상태가 아님을 의미한다. 반면에, 기존 트랙에 할

당될 수 없는 새로운 측정치(음조 검출)이 나타나면 새로운 트랙이 시작된다.

[0093] (34)에서, 공간 큐는 M개의 주파수 변환된 프레임으로부터 추출된다. (31)과 같이, 시간에서의 상관 관계를 이용한 분석을 위해 최근 L 벡터가 저장되었다. 그 결과 TFC(time-frequency-Cue) 맵이 생성되고, 이는 M개의 마이크로폰 각각에 대해  $L \times K \times P$  (여기서,  $P=M-1$ ) 크기의 3차원 배열이다. TFC는 도 5에서 설명된다.

[0094] (35)에서, TFC의 각각의 주파수 성분의 공간 큐가 세그먼트화된다. 아이디어는 L 프레임을 따라 주파수 성분이 다른 화자에서 발원(originate)될 수 있으며, 이것은 공간 큐를 비교하여 관찰될 수 있다는 것이다. 그러나, 단일 프레임 시간  $l$  에서, 주파수 성분은 W-DO 가정으로 인해 단일 화자에서 발원한다고 가정된다. 세그먼트화는 KNN(K nearest neighbors)과 같은 클러스터링에 사용되는 문헌에서 알려진 방법을 사용하여 수행될 수 있다. 클러스터링은 A의 각각의 셀에 인덱스  $c(k, l) \in \mathbb{N}$  를 할당하며, 이는 셀(k,l)이 속한 클러스터를 나타낸다.

[0095] (36)에서, 신호의 주파수 성분은 그룹화되어 각각의 주파수 성분이 EKF에 의해 추적되고 그 신뢰도에 의해 활성화되는 음조의 리스트의 특정 음조에 할당된다. 이것은 음조 중 하나에 할당된 시간 주파수 맵의 k번째 라인(도 4 참조)과 시간 주파수 맵의 다른 라인의 특정 클러스터 인덱스  $c_0(j, l)$  를 가진 모든 값 사이의 샘플 교차 상관(sample-cross-correlation)을 계산하여 수행된다. 이것은 모든 클러스터 인덱스에 대해 수행된다. 샘플 교차 상관은 다음과 같이 주어진다:

$$R(k, j, c_0) = \frac{1}{L} \sum_{\substack{l=0 \\ c(j, l)=c_0}}^{L-1} A(k, l) \cdot A(j, l) \quad (6)$$

[0096] 여기서 A는 시간 주파수 맵이고, k는 음조 중 하나에 속하는 라인의 인덱스이고, j는 A의 임의의 다른 라인이고, L은 A의 열 수이다. 다른 라인의 각각의 음조와 각각의 클러스터 사이의 샘플 교차 상관을 계산한 후, 가장 높은 교차 상관을 갖는 라인  $j_1$ 의 클러스터  $c_1$ 이 개개의 음조로 그룹화되고, 그런 다음 두 번째로 높은 교차 상관을 갖는 라인  $j_2$ 의 클러스터  $c_2$ 가 개개의 음조로 그룹화되는 식이다. 이 프로세스는 샘플 교차 상관이 0.5 x(단일 주파수에서 신호의 평균 에너지)와 같이 적응적으로 설정할 수 있는 임계값  $\kappa$  아래로 떨어질 때까지 반복된다. (35)의 결과는 각각의 음조 주파수가 부여된 주파수 그룹의 세트이다.

[0098] 도 4는 시간 주파수 맵에 대한 음조 검출의 예를 설명한다. (41)은 파라미터  $\ell$  로 표시되는 시간 축이고, (42)는 파라미터  $k$  로 표시되는 주파수 축이다. 이 2차원 배열의 각각의 열은 시간  $\ell$  에서 M개의 주파수 변환된 버퍼의 절대값을 평균한 후 (31)에서 추출된 K-길이 실수 값 벡터이다. 시간에 따른 상관 분석을 위해, L개의 최근 벡터가  $K \times L$  크기의 2차원 배열에 저장된다. (43)에서 두 개의 음조는 상이한 방향의 대각선으로 표시된다.  $k=4, 6, 8$ 에서 고조파를 갖는 음조  $k=2$ 는 4차 고조파가 있기 때문에 신뢰도  $R=20$ 을 가지며,  $k=6, 9$ 에서 고조파를 갖는  $k=3$ 에서 음조는 신뢰도  $R=10$ 을 갖는다. (44)에서  $k=3$  음조는 비활성화되고  $k=2$ 만 활성화된다. 그러나,  $k=2$  음조의 신뢰도는 4차 고조파가 검출되지 않기 때문에  $R=10$ 으로 감소된다(임계값  $\mu$  미만). (45)에서  $k=3$ 의 음조는 다시 활성화되고  $k=2$ 는 비활성화된다. (46)에서  $k=4$ 에서 새로운 음조 후보가 나타나지만 2차 고조파만 검출된다. 따라서 음조가 검출되지 않는다. (47)에서는  $k=3$  음조가 비활성화되어 음조가 검출되지 않는다.

[0099] 도 5는 TFC-맵을 설명하고, 이의 축은 프레임 인덱스(시간)(51), 주파수 성분(52) 및 공간 큐(53)이고, 이는 예를 들어, 각각의 주파수 성분이 도달하는 방향(방위각 및 양각(elevation)) 및 성분의 강도를 표현하는 복소수 값일 수 있다. 인덱스  $\ell$  를 갖는 프레임을 처리하여 주파수 도메인으로 전달될 때, 각각의 주파수 엘리먼트

$$\{k\}_{k=0}^{K-1}$$

에 대해 M개의 복소수의 벡터가 수신된다. 각각의 벡터에 대하여, 최대 M-1개의 공간 큐가 추출된다. 각각의 주파수 성분의 방향 및 강도의 예에서, 이것은 MUSIC 또는 ESPRIT와 같이 당업계에 알려진 어레이 처리를 위한 임의의 방향 찾기 알고리즘(direction-finding algorithm)을 사용하여 수행될 수 있다. 이 알고리즘의 결과는 3차원 공간에서 최대 M-1 방향의 세트이며, 각각은 두 개의 각도와 도달 신호의 추정된 세기

$$\mathbf{p}_p(\ell, k) \triangleq (a(\ell, k), \theta(\ell, k), \phi(\ell, k))_p, \quad p = 1, \dots, P \leq M-1$$

로 표현된다. 큐는 TFC 맵에 정렬되어 셀에서

$\mathbf{p}_{p_0}(\ell_0, k_0)$  는  $\ell_0, k_0, p_0$  에 의해 인덱싱된다.

[0100]

부록

[0101]

음성 증강 모듈의 성능은 원하는 음성 신호만 남기고 모든 간섭 신호를 필터링하는 능력에 의존한다. 간섭 신호는 예를 들어, 다른 화자, 대기 상태의 노이즈, 음악, 모터 노이즈(예를 들어, 자동차 또는 비행기) 및 '칵테일 파티 노이즈(cocktail party noise)'으로 알려진 대규모 군중 노이즈일 수 있다. 음성 증강 모듈의 성능은 일반적으로 SNR(speech-to-noise-ratio) 또는 SIR(speech-to-interference-ratio)를 증강시키는 능력으로 측정되며, 이는 노이즈 및 기타 간섭 신호의 총 파워에 대한 원하는 음성 신호의 파워의 비율 (종종 dB 스케일)로 반영된다.

[0102]

취득 모듈(acquisition module)이 단일 마이크로폰을 포함하는 경우, 이 방법은 단일 마이크로폰 음성 증강이라고 하며, 단일 채널 스펙트럼 빼기, MVDR(Minimum Variance Distortionless Response) 및 반향 제거(echo-cancellation)를 사용한 스펙트럼 추정과 같은 시간 주파수 도메인에서 신호 자체의 통계적 피처를 기반으로 하는 경우가 많다. 하나 이상의 마이크로폰이 사용되는 경우 취득 모듈은 종종 마이크로폰 어레이라고 하며, 방법은 다중 마이크로폰 음성 증강(multi-microphone speech enhancement)이다. 이러한 방법 중 다수는 마이크로폰에 의해 동시에 캡처된 신호 간의 차이를 이용한다. 잘 정립된 방법은 각각의 신호에 가중치 인자를 곱한 후 마이크로폰으로부터의 신호를 합산하는 빔포밍(beamforming)이다. 가중치 인자의 목적은 관심 신호를 조절하기 위해 간섭 신호를 평균화하는 것이다.

[0103]

다시 말해서, 빔포밍은 공간의 주어진 위치에서 방출되는 신호(원하는 화자로부터 원하는 신호)의 파워를 알고리즘적으로 증가시키고 공간의 다른 위치(다른 소스의 간섭 신호)에서 방출되는 신호의 파워를 감소시켜서 빔형성기 출력에서 SIR을 증가시키는 공간 필터를 생성하는 방법이다.

[0104]

DSB(Delay-and-sum beamformer)는 DSB의 가중치 인자를 사용하는 것을 수반하며, 원하는 신호가 그 소스로부터 어레이의 마이크로폰 각각으로 이동하는 상이한 방식에 의해 암시되는 카운터 지연(delay)으로 구성된다. DSB는 자유장 환경(free-field environment)과 같이 각각 단일 방향에서 오는 신호로 제한된다. 결과적으로, 동일한 소스로부터의 신호가 상이한 방식으로 마이크로폰으로 이동하고 복수의 방향으로부터 마이크로폰에 도달하는 잔향 환경에서, DSB 성능은 전형적으로 불충분하다.

[0105]

잔향 환경에서, DSB의 단점을 완화하기 위해, 빔 형성기는 각각의 주파수 성분이 주어진 소스로부터 특정 마이크로폰에 도달하는 방향(방위각 및 양각)을 나타내는 보다 복잡한 음향 전달 함수(ATF : acoustic transfer function)를 사용할 수 있다. DSB 및 기타 DOA 기반 방법에서 가정되는 단일 도달 방향(DOA : direction of arrival)은 동일한 음성 신호의 성분이 다른 방향에서 도달하는 잔향 환경에서는 종종 사실이 아니다. 이것은 벽, 가구 및 사람과 같은 잔향 환경에서 물리적 엘리먼트의 주파수 응답이 상이하기 때문이다. 주파수 도메인의 ATF는 나이퀴스트 대역폭(Nyquist bandwidth)의 각각의 주파수에 복소수를 할당하는 벡터이다. 절대값은 이 주파수와 관련된 경로의 이득을 나타내며, 위상은 경로를 따라 주파수 성분에 추가된 위상을 나타낸다.

[0106]

공간의 주어진 지점과 주어진 마이크로폰 사이의 ATF를 추정하는 것은 주어진 지점에 위치된 라우드스피커(loudspeaker)를 사용하고 알려진 신호를 방출함으로써 수행될 수 있다. 화자의 입력으로부터의 신호와 마이크로폰의 출력 신호를 동시에 취하면 ATF를 쉽게 추정할 수 있다. 라우드스피커는 시스템이 작동하는 동안 사람 화자가 상주할 수 있는 하나 이상의 위치에 위치될 수 있다. 이 방법은 공간의 각각의 지점에 대하여 또는 보다 실질적으로 그리드(grid)의 각각의 지점에 대해 ATF 맵을 생성한다. 그리드에 포함되지 않은 지점의 ATF는 보간법을 사용하여 근사화된다. 그럼에도 불구하고 - 이 방법은 큰 단점이 있다. 첫째, 이 방법을 비실용적으로 만드는 각각의 설치에 대해 시스템을 캘리브레이션해야 한다. 둘째, 사람의 화자와 전자 스피커의 음향 차이로 측정된 ATF와 실제의 ATF가 다르다. 셋째, 특히 또한 화자의 방향을 고려할 때 엄청난 수의 ATF를 측정하는 복잡성 및 넷째, 환경의 변화로 인해 발생할 수 있는 오류이다.

[0107]

ATF에 대한 보다 실용적인 대안은 실제 애플리케이션에서 ATF 추정 방법의 단점에 대한 해결책으로서 상대 전달 함수(RTF : relative transfer function)이다. RTF는 주어진 소스와 어레이의 두 마이크로폰 사이의 ATF 차이로, 이는 주파수 도메인에서 두 ATF의 스펙트럼 표현 간의 비율의 형태를 취한다. ATF와 마찬가지로, 주파수 도

메인의 RTF는 각각의 주파수에 복소수를 할당한다. 절대값은 두 마이크로폰 사이의 게인 차이로, 이는 마이크로폰이 서로 가까울 때 종종 1에 가깝고, 일부 조건에서 위상은 소스의 입사각을 반영한다.

[0108] TF-LCMV(Transfer function based linear constrained minimum variance) 빔 형성기는 다수의 마이크로폰 애플리케이션에서, 출력 신호의 음성 성분이 마이크로폰 신호 중 하나의 음성 성분과 동일하다는 제약 조건에 따라 출력 에너지를 최소화함으로써 음성 왜곡을 제한하면서 노이즈를 줄일 수 있다.  $N = N_d + N_i$  소스가 주어지면,  $N_d$  원하는 음성 소스를 추출하는 문제,  $N_i$  간섭 소스에 의한 오염, 및 정적 노이즈를 고려한다. 수신된 신호 각각은 마이크로폰을 포함하는 임의의 어레이에 의해 선택되기 전에 음향 매체를 통해 전파된다. 각각의 마이크로폰의 신호는 길이의 프레임으로 세그먼트화되고 각각의 프레임에 FFT가 적용된다. 주파수 도메인에서,  $m$ 번째 마이크로폰과  $n$ 번째 소스의  $\ell$ -번째 프레임의  $k$ 번째 주파수 성분을 각각,  $z_m(\ell, k) \in \mathbb{C}$  및  $s_n(\ell, k) \in \mathbb{C}$  로 표시하자. 유사하게,  $n$ 번째 소스와  $m$ 번째 마이크로폰 사이의 ATF는  $g_{m,n}(\ell, k)$  이고,  $m$ 번째 마이크로폰에서의 노이즈는  $v_m(\ell, k)$  이다. 행렬 형태로 수신된 신호는 다음과 같이 주어진다 :

[0109] 
$$\mathbf{z}(\ell, k) = \mathbf{G}(\ell, k) \mathbf{s}(\ell, k) + \mathbf{v}(\ell, k) \in \mathbb{C}^M \quad (7)$$

[0110] 여기서,  $\mathbf{z}(\ell, k) = [z_1(\ell, k), \dots, z_M(\ell, k)]^T \in \mathbb{C}^M$  는 센서 벡터,  $\mathbf{s}(\ell, k) = [s_1(\ell, k), \dots, s_N(\ell, k)]^T \in \mathbb{C}^N$  는 소스 벡터,  $\mathbf{G}(\ell, k) \in \mathbb{C}^{M \times N}$  는 ATF 행렬이어서,  $[\mathbf{G}(\ell, k)]_{m,n} = g_{m,n}(\ell, k) \in \mathbb{C}$  , 및  $\mathbf{v}(\ell, k) = [v_1(\ell, k), \dots, v_M(\ell, k)]^T \in \mathbb{C}^M$  는 소스 중 어느 것과도 상관되지 않는 가산 정적 노이즈(additive stationary noise)이다. 동등하게, (7)은 RTF를 사용하여 공식화될 수 있다. 일반성의 손실 없이,  $n$ 번째 음성 소스의 RTF  $h_{m,n}(\ell, k) \in \mathbb{C}$  는  $m$ 번째 마이크로폰에서  $n$ 번째 음성 성분과 첫 번째 마이크로폰에서 각각의 성분 사이의 비율로 정의될 수 있다 즉,  $h_{m,n}(\ell, k) = g_{m,n}(\ell, k) / g_{1,n}(\ell, k)$  . (7)의 신호는 RTF 행렬을  $\mathbf{H}(\ell, k) \in \mathbb{C}^{M \times N}$  사용하여 공식화될 수 있어서  $[\mathbf{H}(\ell, k)]_{m,n} = h_{m,n}(\ell, k) \in \mathbb{C}$  , 벡터 표기법으로 :

[0111] 
$$\mathbf{z}(\ell, k) = \mathbf{H}(\ell, k) \mathbf{x}(\ell, k) + \mathbf{v}(\ell, k) \in \mathbb{C}^M \quad (8)$$

[0112] 여기서,  $x_n(\ell, k) = g_{1,n}(\ell, k) s_n(\ell, k) \in \mathbb{C}$  는 변경된 소스 신호이다.

[0113] 어레이 측정  $\mathbf{z}(\ell, k)$  이 주어지면  $N_d$  원하는 소스의 혼합(mixture)을 추정할 필요가 있다. 마이크로폰 신호  $\mathbf{y}(\ell, k) = \mathbf{w}^H(\ell, k) \mathbf{z}(\ell, k) \in \mathbb{C}$  에 빔 형성기  $\mathbf{w}(\ell, k) \in \mathbb{C}^M$  를 적용하여 원하는 신호의 추출이 달성될 수 있다.  $M \geq N$  이라고 가정하면,  $\mathbf{w}(\ell, k) \in \mathbb{C}^M$  은 LCMV 기준을 충족하도록 선택할 수 있다:

[0114] 
$$\mathbf{w}(\ell, k) = \arg \min_{\mathbf{w}} \{ \mathbf{w}^H(\ell, k) \Phi_{\mathbf{v}}(\ell, k) \mathbf{w}(\ell, k) \} \quad \text{s.t.} \quad \mathbf{H}^H(\ell, k) \mathbf{w}(\ell, k) = \mathbf{c}(\ell, k) \quad (9)$$

[0115] 여기서,  $\Phi_{\mathbf{v}}(\ell, k) \in \mathbb{C}^{M \times M}$  는  $\mathbf{v}(\ell, k)$  의 파워 스펙트럼 밀도(PSD : power spectral density) 행렬이고,  $\mathbf{c}(\ell, k) \in \mathbb{C}^{N \times 1}$  는 제약 벡터이다.

[0116] (9)에 대한 가능한 솔루션은 다음과 같다 :

[0117] 
$$\mathbf{w}_{\text{LCMV}}(\ell, k) = \Phi_{\mathbf{w}}^{-1}(\ell, k) \mathbf{H}(\ell, k) \left( \mathbf{H}^H(\ell, k) \Phi_{\mathbf{w}}^{-1}(\ell, k) \mathbf{H}(\ell, k) \right)^{-1} \mathbf{c}(\ell, k) \quad (10)$$

[0118] (7) 및 (8) 및 제약 세트에 기초하여, 빔 형성기 출력에서 원하는 신호의 성분은  $d(\ell, k) = \mathbf{c}^H(\ell, k) \mathbf{x}(\ell, k) \in \mathbb{C}$ 로 주어지고, 즉, 빔 형성기의 출력은 제1(기준) 마이크에 의해 측정된 원하는 신호의 성분의 혼합이다.

[0119] RTF의  $\ell$ -번째 세트로부터 및 각각의 주파수 성분  $k$ 에 대해, 입사각  $\theta_p(\ell, k)$ ,  $p = 1, \dots, P \leq M-1$ , 및 양각  $\phi_p(\ell, k)$ 을 갖는 최대  $M-1$  소스까지의 세트는 기준으로 정의된 마이크로폰 중 하나에서 취해진 강도  $a_p(\ell, k)$ 와 함께, 예를 들어, 위상차 기반 알고리즘을 사용하여 추출될 수 있다. 이러한 3-튜플  $\mathbf{p}_p(\ell, k) \triangleq (a(\ell, k), \theta(\ell, k), \phi(\ell, k))_p \in \mathbb{R}^3$ 은 종종 공간 큐라고 한다.

[0120] TF-LCMV는 잔향 환경에서 상이한 위치로부터의  $M$  센서를 포함하는 어레이에 충돌하는  $M-1$  음성 소스를 추출하기 위한 적용 가능한 방법이다. 그러나 TF-LCMV가 작동하기 위한 필요 조건은 열(column)이 환경의 모든 활성 소스

$\mathbf{H}(\ell, k)$ 의 RTF 벡터인 RTF 행렬이 알려져 있고 TF-LCMV에서 사용할 수 있다는 것이다. 이를 위해서는 각각의 주파수 성분을 소스 화자에 연관시켜야 한다.

[0121] 보충 정보 없이 신호에 소스를 할당하기 위해 여러 방법이 사용될 수 있다. 주요 방법군은 BSS(Blind Source Separation)라고 하며, 이는 관찰된 혼합에서 알려지지 않은 신호 또는 소스를 복구한다. 주파수 도메인에서 BSS의 주요 약점은 각각의 주파수에서, 혼합 행렬의 열 벡터(BSS에 의해 추정됨)가 무작위로 순열(permute)되며, 이 무작위 순열에 대한 지식 없이는, 주파수에 걸쳐 결과를 조합하는 것이 개시된 바와 같이 어렵다는 것이다.

[0122] BSS는 음조 정보에 의해 도움을 받을 수 있다. 단, 화자의 성별은 우선적으로 요구된다.

[0123] BSS는 주파수 도메인에서 사용될 수 있으며, 혼합 행렬의 특정 열을 벡터의 최대 엘리먼트에 해당하는 소스에 할당하는 최대 크기 방법을 사용하여 추정된 혼합 행렬의 모호성을 해결할 수 있다. 그럼에도 불구하고 - 이 방법은 각각의 주파수에서 가장 강한 성분이 실제로 가장 강한 소스에 속한다고 가정하기 때문에 소스의 스펙트럼 분포에 크게 의존한다. 그러나 다른 화자가 다른 주파수에서 강도 피크를 유발할 수 있으므로 이 조건은 자주 충족되지 않는다. 대안적으로, 특정 시간의 활성 소스에 대한 정보가 혼합 행렬의 모호성을 해결하는 데 사용되도록 VAD(voice activity detection)라고도 알려진 소스 활성 검출이 사용될 수 있다. VAD의 단점은 특히 다수의 화자 환경에서 음성 일시 중지(voice-pause)를 강력하게 검출할 수 없다는 것이다. 또한 이 방법은 한 번에 한 명 초과 화자가 대화에 참여하지 않는 경우에만 효과적이며 상대적으로 긴 트레이닝 기간(training period)이 필요하며 이 기간 동안 움직임에 민감하다.

[0124] TF-LCMV 빔 형성기는 바이노럴(binaural) 큐 생성기와 함께 바이노럴 음성 증강 시스템을 위한 확장 버전과 함께 사용될 수 있다. 음향 큐는 입력 신호의 노이즈 성분에서 음성 성분을 분리하는 데 사용된다. 이 기술은 "각 테일 파티" 환경에서 고유한 음성 소스로부터의 신호를 클러스터링하기 위해 고유한 지각 신호의 사용을 제안하는 청각 장면 분석 이론<sup>1</sup>을 기반으로 한다. 음성 분리에 사용될 수 있는 프리미티브(primitive) 그룹화 큐의 예에는 주파수 대역에 걸친 공통 온셋(onset)/오프셋(offset), 음조(기본 주파수), 공간의 동일한 위치, 시간 및 스펙트럼 변조, 음조 및 에너지 연속성 및 부드러움이 포함된다. 그러나 이 방법의 기본 가정은 원하는 음성 신호의 모든 성분이 거의 동일한 방향을 갖는다는 것이다. 즉, 머리 관련 전달 함수를 사용하여 보상하는 것이 제안된 머리 그림자 효과(head-shadow effect)의 효과를 제외한 거의 자유장 조건(free-field condition)이다. 잔향 환경(reverberant environment)에서는 발생하지 않을 것이다.

[0125] 다수의 화자가 동시에 활성화된 경우에도 화자의 스펙트럼 콘텐츠는 대부분의 시간 주파수 지점에서 중첩되지 않는다는 점에 유의해야 한다. 이를 W-서로소 직교성(Disjoint Orthogonality) 또는 간단히 W-DO라고 한다. 이것은 시간 주파수 도메인에서 음성 신호의 희소성(sparseness)에 의해 정당화될 수 있다. 이 희소성에 따르면, 특정 시간 주파수 지점에서 두 화자가 동시에 활성화될 확률은 매우 낮다. 즉, 다수의 동시 화자의 경우, 각각의 시간 주파수 지점은 화자 중 하나의 스펙트럼 콘텐츠에 해당할 가능성이 가장 높다.

- [0126] W-D0는 어느 정도 W-D0인 신호의 특정 클래스를 정의함으로써 BSS를 가능하게 하는 데 사용될 수 있다. 이것은 계산적으로 경제적인 1차 통계만 사용할 수 있다. 또한, 소스가 W-D0이고 동일한 공간 위치를 점유하지 않는 경우 두 개의 마이크로폰만 사용하여 임의의 수의 신호 소스를 디믹스(de-mix)할 수 있다. 그러나 이 방법은 모든 주파수에서 동일한 기본 혼합 행렬을 가정한다. 이 가정은 상이한 주파수에 걸쳐 추정된 혼합 계수의 히스토그램(histogram)을 사용하는 데 필수적이다. 그러나 이 가정은 종종 잔향 환경에서는 적용되지 않고 자유장에서만 적용된다. 다수의 경로의 경우에 이 방법의 확장은 다수의 경로에서 무시할 수 있는 에너지로 제한되거나 히스토그램이 번지지만 단일 피크를 유지하도록 충분히 부드러운 컨볼루션 혼합 필터(convolutive mixing filter)로 제한된다. 이 가정은 상이한 경로 간의 차이가 종종 너무 커서 부드러운 히스토그램을 생성할 수 없는 잔향 환경에서는 적용되지 않는다.
- [0127] 제안된 솔루션은 잔향 환경에서 수행되고 불필요한 가정 및 제약에 의존할 필요가 없다는 것이 발견되었다. 솔루션은 사전 정보 없이도 심지어 큰 트레이닝 과정 없이도 각각의 주파수에서 주어진 소스의 감쇠 및 지연 추정을 감쇠 지연 공간의 단일 지점으로 제한하지 않고, 단일 클러스터를 생성하기 위해 단일 소스의 감쇠 지연 값의 추정 값을 제한하지 않고, 혼합 사운드의 수를 2개로 제한하지 않아도 작동할 수 있다.
- [0128] 음성 인식 엔진에 대한 소스 분리
- [0129] 음성 사용자 인터페이스(VUI : voice user interface)는 사람의 화자와 기계 간의 인터페이스이다. VUI는 하나 이상의 마이크로폰을 사용하여 음성 신호를 수신하고 종종 음성 신호를 텍스트로 전사(transcribe)하여 음성 신호를 디지털 시그니처(signature)으로 변환하고, 이는 화자의 의도를 추론하는 데 사용된다. 그런 다음 기계는 기계가 디자인된 애플리케이션에 따라 화자의 의도에 응답할 수 있다.
- [0130] VUI의 핵심 컴포넌트는 디지털화된 음성 신호를 텍스트로 변환하는 자동 음성 인식 엔진(ASR : automatic speech recognition)이다. ASR의 성능, 즉 텍스트가 음향 음성 신호를 얼마나 정확하게 설명하느냐는 입력 신호를 ASR의 요구 사항과 일치시키는 데 크게 좌우된다. 따라서 VUI의 다른 컴포넌트는 획득한 음성 신호를 ASR에 공급하기 전에 증강시키도록 디자인되었다. 이러한 컴포넌트는 노이즈 억제, 반향 제거 및 소스 분리 등을 들 수 있다.
- [0131] 음성 증강의 중요한 컴포넌트 중 하나는 여러 소스로부터 도달하는 음성 신호를 분리하기 위한 소스 분리(SS : source separation)이다. 두 개 이상의 마이크로폰 어레이를 가정할 때, 각각의 마이크로폰에 의해 획득된 신호는 환경의 모든 음성 신호와 노이즈 및 음악과 같은 다른 간섭의 혼합이다. SS 알고리즘은 모든 마이크로폰으로부터 혼합된 신호를 가져와 그것들을 컴포넌트로 분해한다. 즉, 소스 분리의 출력은 신호 세트이며, 각각은 특정 화자의 음성 신호인 특정 소스의 신호, 음악 또는 노이즈를 나타낸다.
- [0132] 소스 분리를 개선할 필요성이 증가하고 있다.
- [0133] 도 6은 오프라인 트레이닝에서 음성 인식 체인의 예를 예시한다. 체인은 디지털화된 음향 신호 세트를 제공하는 마이크로폰(511) 어레이를 종종 포함한다. 디지털화된 음향 신호의 수는 어레이(511)를 구성하는 마이크로폰의 수와 동일하다. 각각의 디지털화된 음향 신호는 마이크로폰(511)의 어레이 부근에 있는 모든 음향 소스의 혼합을 포함하며, 이는 사람 화자, TV와 같은 합성 화자, 음악과 노이즈이다. 디지털화된 음향 신호는 전처리(pre-processing) 스테이지(512)로 전달된다. 전처리 단계(512)의 목적은 에코, 잔향 및 노이즈와 같은 간섭을 제거함으로써 디지털화된 음향 신호의 품질을 개선하는 것이다. 전처리 단계(512)는 일반적으로 디지털화된 음향 신호들 사이의 통계적 연관을 사용하는 멀티채널 알고리즘을 사용하여 수행된다. 전처리 단계(512)의 출력은 처리된 신호 세트이며, 일반적으로 이 스테이지에 대한 입력에서 디지털화된 음향 신호의 수와 동일한 수의 신호를 갖는다. 처리된 신호 세트는 마이크로폰 어레이 부근의 각각의 소스로부터 음향 신호를 추출하는 것을 목표로 하는 소스 분리(SS : source separation) 스테이지(513)로 포워딩된다. 다시 말해서, SS 스테이지(513)는 신호 세트를 취하고, 각각은 상이한 소스로부터 수신된 음향 신호의 상이한 혼합이며, 각각의 신호가 주로 단일 특정 소스로부터의 단일 음향 신호를 포함하도록 신호 세트를 생성한다. 음성 신호의 소스 분리는 빔포밍과 같은 소스 배치의 기하학적 고려 사항을 사용하거나 또는 독립 성분 분석과 같은 음성 신호의 특성을 고려하여 수행될 수 있다. 분리된 신호의 수는 일반적으로 마이크로폰 어레이(511) 부근의 활성 소스의 수와 동일하지만, 마이크로폰의 수보다는 적다. 분리된 신호 세트는 소스 선택기(514)로 포워딩된다. 소스 선택기의 목적은 음성 신호가 인식되어야 하는 관련 음성 소스를 선택하는 것이다. 소스 선택기(514)는 미리 정의된 트리거 단어를 발음하는 소스가 선택되도록 트리거 단어 검출기를 사용할 수 있다. 대안적으로, 소스 선택기(514)는 마이크로폰(511)의 어레이에 대한 미리 정의된 방향과 같이 마이크로폰(511)의 어레이 부근에서 소스의 위치를 고려할 수 있다. 또한, 소스 선택기(514)는 음성 신호의 미리 정의된 음향 시그니처를 사용하여 이 시그니처와 일치하는 소스를 선

택할 수 있다. 소스 선택기(514)의 출력은 음성 인식 엔진(515)으로 포워딩되는 단일 음성 신호이다. 음성 인식 엔진(515)은 디지털화된 음성 신호를 텍스트로 변환한다. 이 기술 분야에 알려진 음성 인식을 위한 많은 방법이 있으며, 이들 대부분은 음성 신호로부터 피처를 추출하고 이러한 피처를 미리 정의된 어휘와 비교하는 것을 기반으로 한다. 음성 인식 엔진(515)의 주요 출력은 입력 음성 신호와 연관된 텍스트 스트링(516)이다. 미리 정의된 텍스트(518)는 오프라인 트레이닝에서 마이크로폰에 받음된다. 오류 (519)는 ASR (516)의 출력을 이 텍스트와 비교하여 계산된다. 비교(517)는 간단한 단어 카운팅 또는 단어의 의미를 고려하고 상이한 단어의 오검출에 적절하게 가중치를 주는 보다 정교한 비교 방법을 사용하여 수행될 수 있다. 오류(519)는 그런 다음 오류를 최소화하는 값을 찾기 위해 파라미터 세트를 수정하기 위해 SS(513)에 의해 사용된다. 이것은 최소 제곱(least square), 확률적 기울기, 신경망(NN : neural network) 및 그 변형과 같은 감독된 추정 또는 최적화 방법으로 수행될 수 있다.

[0134] 도 7은 실시간 트레이닝, 즉 시스템의 정상 작동 동안의 음성 인식 체인의 예를 예시한다. VUI 작동 중 사람 화자가 발음한 실제 텍스트를 알 수 없으며 감독된 오류(519)는 사용할 수 없다. 대안은 실제 음성 텍스트에 대한 기준이 없을 때 실시간 애플리케이션을 위해 개발된 신뢰 점수(confidence score)(521)이며 애플리케이션은 ASR 출력의 신뢰도 수준을 알면 이점이 있다. 예를 들어, 신뢰도 점수가 낮을 때 시스템은 사용자와 보다 직접적인 대화가 수행되는 적절한 분기로 이동할 수 있다. 신뢰도 점수를 추정하는 많은 방법이 있으며, 이들 대부분은 음성 텍스트를 알고 있을 때 계산될 수 있는 오류와 높은 상관 관계를 목표로 한다. 실시간 트레이닝에서, 신뢰도 점수(521)는 오류 추정기(522)에 의해 감독된 오류 (supervised error)(519)로 변환된다. 신뢰도 점수가 이론적인 감독된 오류와 높은 상관 관계가 있는 경우, 오류 추정기(522)는 단순 측 변환일 수 있다. 신뢰도 점수(521)의 범위는 0에서 100 사이이며 목표는 더 높은 값을 얻는 것이지만, 감독된 오류는 0에서 100까지이고 목표는 더 낮은 값을 얻는 것이다.  $estimated\_error = 100 - confidence\_score$  형식의 단순 측 변환이 오류 추정기(522)로 사용될 수 있다. 추정된 오류(519)는 오프라인 트레이닝과 마찬가지로 SS의 파라미터를 트레이닝하는데 사용될 수 있다.

[0135] 도 8은 전형적인 SS(513)의 트레이닝 메커니즘을 예시한다. 소스 분리기(SS)(513)는 전처리 스테이지(512)로부터 혼합 신호 세트를 수신하고 분리된 신호를 소스 선택기(514)에 공급한다. 일반적으로 음향 신호, 특히 음성 신호의 소스 분리는 주파수 도메인에서 수행된다. 전처리 단계(512)로부터의 혼합 신호는 먼저 주파수 도메인(553)으로 변환된다. 이것은 혼합 신호를 동일한 길이의 세그먼트로 분할함으로써 수행되고, 후속 세그먼트 사이에는 중첩 기간이 있다. 예를 들어, 세그먼트의 길이가 1024 샘플로 결정되고 중첩 기간이 25%인 경우, 혼합 신호 각각은 각각 1024 샘플의 세그먼트로 분할된다. 상이한 혼합 신호로부터의 동시 세그먼트 세트를 일괄처리 묶음(batch)라고 한다. 세그먼트의 각각의 일괄처리 묶음은 이전 일괄처리 묶음 이후에 768개 샘플을 시작한다. 혼합 신호 세트의 세그먼트는 동기화된다 즉, 동일한 일괄처리 묶음에 속하는 모든 세그먼트의 시작점이 동일하다는 것에 유의한다. 일괄처리 묶음의 세그먼트 길이와 중첩 기간은 모델 파라미터(552)로부터 가져온다.

[0136] 디믹싱 알고리즘(demixing algorithm)(554)은 주파수 변환(553)으로부터 도달된 세그먼트의 일괄처리 묶음을 분리한다. 많은 다른 알고리즘과 마찬가지로, 소스 분리(SS) 알고리즘은 모델 파라미터(552) 세트와 함께 수학적 모델 세트를 포함한다. 수학적 모델은 SS가 물리적 현상에 대처하는 방식과 같은 작동 방법 예를 들어, 다중 경로(multipath)를 수립한다. 모델 파라미터(552) 세트는 소스 신호의 특정 피처, 이러한 신호를 수신하는 자동 음성 인식 엔진(ASR)의 아키텍처, 환경의 기하학적 구조, 심지어 사람 화자에 대한 SS의 동작을 조정한다.

[0137] 세그먼트의 디믹싱된 일괄처리 묶음은 역 주파수 변환(555)으로 포워딩되어 시간 도메인으로 다시 변환된다. 주파수 변환 스테이지(553)에서 사용된 동일한 세트의 모델 파라미터(552)가 역 주파수 변환 스테이지(555)에서 사용된다. 예를 들어, 중첩 기간은 후속 일괄처리 묶음으로부터 시간 도메인의 출력 신호를 재구성하는 데 사용된다. 이것은 예를 들어, 역 주파수 변환 후에 결과 출력 신호가 중첩 영역에 걸쳐 0에서 1 사이의 범위인 적절한 가중 함수를 사용하여 중첩 간격을 중첩 및 추가하여 재구성되는 중첩 추가 방법(overlap-add method)을 사용하여 수행되어서 전체 에너지가 보존된다. 다시 말해, 전자의 일괄처리 묶음의 중첩 세그먼트는 페이드 아웃(fade out)되고 후자의 오버랩 세그먼트는 페이드 인(fade in)된다. 역 주파수 변환 블록의 출력은 소스 선택기(514)로 포워딩된다.

[0138] 모델 파라미터(552)는 주파수 변환 블록(553), 디믹싱 블록(554) 및 역 주파수 변환 블록(555)에 의해 사용되는 파라미터 세트이다. 주파수 변환(553)에 의해 수행되는 동일한 길이의 세그먼트로의 혼합된 신호의 분할은 실시간 클록과 같은 클록 메커니즘에 의해 진행된다. 각각의 페이스(pace)에서, 주파수 변환 블록(553), 디믹싱 블록(554) 및 역 주파수 변환 블록(555) 각각은 모델 파라미터(552)로부터 파라미터를 추출한다. 그런 다음 이러한 파라미터는 주파수 변환 블록(553), 디믹싱 블록(554) 및 역 주파수 변환 블록(555)에서 실행되는 수학적 모

텔에서 대체된다.

- [0139] 정정기(corrector)(551)는 오류 추정기로부터 오류(519)를 감소시키는 것을 목표로 하여 모델 파라미터(552)의 세트를 최적화한다. 정정기(551)는 오류(519) 및 모델 파라미터(552)의 현재 세트를 수신하고 정정된 모델 파라미터(552) 세트를 출력한다. 파라미터 세트의 정정은 사전에(오프라인) 또는 VUI 동작 동안에(실시간) 수행될 수 있다. 오프라인 트레이닝에서, 모델 파라미터(552) 세트를 수정하는 데 사용되는 오류(519)는 마이크로폰에 발음되고 ASR의 출력을 이 텍스트와 비교하는 미리 정의된 텍스트를 사용하여 추출된다. 실시간 트레이닝에서, ASR의 신뢰도 점수로부터 오류(519)가 추출된다.
- [0140] 그런 다음, 오류는 오류를 최소화하는 값을 찾기 위해 파라미터 세트를 수정하는 데 사용된다. 이것은 감독된 추정 또는 최적화 방법, 바람직하게는 골든 섹션 검색(golden section search), 그리드 검색 및 Nelder-Mead와 같은 파생 없는 방법(derivative free method)으로 수행할 수 있다.
- [0141] Nelder-Mead 방법(또한 다운힐 심플렉스 방법(downhill simplex method), 아메바 방법(amoeba method) 또는 폴리토프 방법(polytope method))은 다차원 공간에서 목적 함수의 최소값 또는 최대값을 찾는 데 사용되는 일반적으로 적용되는 수치 방법이다. 직접 검색 방법(함수 비교 기반)이며 도함수를 알 수 없는 비선형 최적화 문제에 자주 적용된다.
- [0142] Nelder-Mead는 여러 파라미터의 함수로서 오류(519)의 로컬 최소값을 반복적으로 찾는다. 이 방법은 심플렉스(simplex) (N 차원의 일반화된 삼각형)를 결정하는 일련의 값으로 시작한다. 로컬 최소값이 심플렉스 내에 존재한다는 가정이 있다. 각각의 반복에서 심플렉스의 꼭짓점(vertex)에서 오류가 계산된다. 최대 오류를 갖는 꼭짓점을 새로운 꼭짓점으로 교체하여 심플렉스의 볼륨을 줄인다. 이것은 심플렉스 볼륨이 미리 정의된 볼륨보다 작고 최적 값이 꼭짓점 중 하나가 될 때까지 반복된다. 이 과정은 정정기(551)에 의해 수행된다.
- [0143] 골든 섹션 검색은 최소값이 존재하는 것으로 알려진 값의 범위를 연속적으로 좁힘으로써 오류(519)의 최소값을 찾는다. 골든 섹션 검색은 파라미터의 함수로서 엄격한 단봉 오류(unimodal error)를 요구한다. 범위를 좁히는 동작은 정정기(551)에 의해 수행된다.
- [0144] 골든 섹션 검색은 극값이 존재하는 것으로 알려진 값의 범위를 연속적으로 좁혀 엄밀한 단봉 함수의 극값(최소 또는 최대)을 찾는 기술이다.(www.wikipedia.org).
- [0145] 그리드 검색은 최적화될 하나 이상의 파라미터와 연관된 값 세트를 통해 반복된다. 둘 이상의 파라미터가 최적화되면, 세트의 각각의 값은 길이가 파라미터 수와 동일한 벡터이다. 각각의 값에 대해, 오류(519)가 계산되고 최소 오류에 해당하는 값이 선택된다. 값 세트를 통한 반복은 정정기(551)에 의해 수행된다.
- [0146] 그리드 검색 - 하이퍼파라미터 최적화를 수행하는 전통적인 방법은 학습 알고리즘의 하이퍼파라미터 공간의 수동으로 지정된 서브 세트를 통한 완전한 검색인 그리드 검색, 또는 파라미터 스위프(parameter sweep)이었다. 그리드 검색 알고리즘은 전형적으로 트레이닝 세트에 대한 교차 유효성 검증(cross-validation) 또는 보류된 유효성 검증 세트에 대한 평가에 의해 측정된 일부 성능 메트릭에 의해 가이드되어야 한다. 기계 학습기(machine learner)의 파라미터 공간에는 특정 파라미터에 대한 실수값 또는 무한값 공간이 포함될 수 있으므로 그리드 검색을 적용하기 전에 수동으로 경계를 설정하고 이산화(discretization)해야 할 수 있다. (www.wikipedia.org).
- [0147] 모든 최적화 방법은 동일한 세트의 분리된 음향 신호를 사용하여 오류(519)의 연속적인 계산을 필요로 한다. 이것은 시간 소모적인 프로세스이므로 연속적으로 수행되지 않을 수 있지만, 오류(519)(연속적으로 계산됨)가 미리 정의된 임계값 즉, 10% 오류를 초과하는 경우에만 수행할 수 있다. 이 경우, 두 가지 접근 방식을 취할 수 있다.
- [0148] 한 가지 접근 방식은 병렬 스레드 또는 다중 코어를 사용하여 시스템의 정상 작동과 병렬로 최적화를 작동하는 것이다. 즉, 시스템의 정상 작동 작업과 병렬로 블록(513, 514, 515, 522)을 수행하는 하나 이상의 병렬 작업이 있다. 병렬 작업에서, 길이가 1-2초인 혼합 신호의 일괄처리 묶음이 전처리(512)에서 취해지며 반복적으로 분리되고(513), 다른 세트의 모델 파라미터(552)로 해석(514, 515)된다. 오류(519)가 각각의 이러한 사이클에 대하여 계산된다. 각각의 사이클에서, 모델 파라미터의 세트는 최적화 방법에 따라 정정기(551)에 의해 선택된다.
- [0149] 두 번째 접근 방식은 룸(room)에 음성이 없을 때 최적화를 작동하는 것이다. 음성 활동 검출(VAD) 알고리즘을 사용하여 사람의 음성이 없는 기간을 검출할 수 있다. 이 기간은 병렬 스레드 또는 다중 코어의 필요성을 제외한, 첫 번째 접근 방식에서와 동일한 방식으로 모델 파라미터(552)를 최적화하는데 사용된다.
- [0150] (552)에서 각각의 파라미터에 대해 적절한 최적화 방법이 선택되어야 한다. 방법 중 일부는 단일 파라미터에 적

용되고 일부는 파라미터 그룹에 적용된다. 다음 텍스트는 음성 인식 성능에 영향을 미치는 몇 가지 파라미터를 제안한다. 또한, 파라미터의 특성에 따라 최적화 방법이 제안된다.

- [0151] 세그먼트 파라미터의 길이.
- [0152] 세그먼트 파라미터의 길이는 FFT/IFFT와 관련된다. 전형적으로, 분리된 음소의 피처를 사용하는 ASR은 20mSec 정도의 짧은 세그먼트를 필요로 하는 반면, 연속적인 음소의 피처를 사용하는 ASR은 100-200mSec 정도의 세그먼트를 사용한다. 한편, 세그먼트의 길이는 룬의 잔향 시간과 같은 시나리오의 영향을 받는다. 세그먼트 길이는 200-500mSec 정도의 잔향 시간 정도이어야 한다. 세그먼트 길이에 대한 스위트 지점(sweet point)이 없으므로, 이 값은 시나리오 및 ASR에 최적화되어야 한다. 전형적인 값은 샘플 기준으로 100-500mSec이다. 예를 들어, 8kHz의 샘플링 속도는 800-4000개 샘플의 세그먼트 길이를 의미한다. 이것은 연속 파라미터이다.
- [0153] 이 파라미터의 최적화는 중복 기간과 함께 골든 섹션 검색 또는 Nelder - Mead와 같은 다양한 최적화 방법을 사용하여 수행될 수 있다. 골든 섹션 검색을 사용할 때, 알고리즘에 대한 입력은 가능한 최소 및 최대 길이 예를 들어, 10mSec 내지 500mSec 및 오류 함수(519)이다. 출력은 오류 함수(519)를 최소화하는 세그먼트의 길이이다. 중첩 기간과 함께 Nelder-Mead를 사용하면, 입력은 세그먼트 길이와 중첩 기간의 세 개의 2-튜플(tuple) 세트 예를 들어, (10mSec, 0%), (500mSec, 10%) 및 (500mSec, 80%) 및 오류 함수(519)이며, 출력은 최적의 세그먼트 길이와 최적의 중첩 기간이다.
- [0154] 중첩 기간(overlap period)
- [0155] 중첩 기간 파라미터는 FFT/IFFT와 관련된다. 중첩 기간은 세그먼트화로 인해 음소가 간과되는 것을 피하기 위해 사용된다. 즉, 후속 세그먼트 사이에서 분할되는 음소이다. 세그먼트의 길이로서, 중첩 기간은 ASR이 사용하는 피처에 따라 다르다. 전형적인 범위 - 세그먼트 길이의 0-90%. 이것은 연속 파라미터이다.
- [0156] 이 파라미터의 최적화는 세그먼트 길이에 대한 골든 섹션 검색(Golden section search) 또는 Nelder-mead와 같은 다양한 최적화 방법을 사용하여 수행될 수 있다. 골든 섹션 검색을 사용할 때, 알고리즘에 대한 입력은 가능한 최소 및 최대 중첩 기간 예를 들어, 0% 내지 90%이고 오류 함수(519)이다. 출력은 오류 함수(519)를 최소화하는 중첩 기간이다.
- [0157] 윈도우(window). 윈도우 파라미터는 FFT/IFFT와 관련된다. 주파수 변환(553)은 종종 세그먼트화의 효과를 완화하기 위해 윈도우(windowing) 사용한다. Kaiser 및 Chebyshev와 같은 일부 윈도우는 파라미터화된다. 이것은 윈도우의 파라미터를 변경하여 윈도우의 효과를 제어할 수 있다는 것을 의미한다. 전형적인 범위는 윈도우 유형에 따라 다르다. 이것은 연속 파라미터이다. 이 파라미터의 최적화는 골든 섹션 검색과 같은 다양한 최적화 방법을 사용하여 수행될 수 있다. 골든 섹션 검색을 사용할 때, 알고리즘에 대한 입력은 윈도우 유형과 오류 함수(519)에 따라 달라지는 윈도우 파라미터의 최소값과 최대값이다. 예를 들어, Kaiser 윈도우의 경우, 최소값과 최대값 (0,30)이다. 출력은 최적의 윈도우 파라미터이다.
- [0158] 샘플링 속도(Sampling rate)
- [0159] 샘플링 속도 파라미터는 FFT/IFFT와 관련된다. 샘플링 속도는 음성 인식 성능에 영향을 미치는 중요한 파라미터 중 하나이다. 예를 들어, 16kHz보다 낮은 샘플링 속도에 대해 좋지 않은 결과를 보여주는 ASR이 있다. 다른 것은 4 또는 8kHz에서도 잘 작동할 수 있다. 전형적으로 이 파라미터는 ASR을 선택할 때 한 번 최적화된다. 전형적인 범위는 4, 8, 16, 44.1, 48kHz이다. 이 파라미터는 이산 파라미터(discrete parameter)이다. 이 파라미터의 최적화는 그리드 검색과 같은 다양한 최적화 방법을 사용하여 수행할 수 있다. 알고리즘에 대한 입력은 그리드 검색이 수행되는 값이다 - 예를 들어, (4, 8, 16, 44.1, 48) kHz의 샘플링 속도와 오류 함수 (519)이다. 출력은 최적의 샘플링 속도이다.
- [0160] 필터링(filtering)
- [0161] 필터링 파라미터는 디믹싱(demixing)과 관련이 있다. 일부 ASR은 제한된 주파수 범위를 나타내는 피처를 사용한다. 따라서, 소스 분리(513) 후에 분리된 신호를 필터링하는 것은 ASR에서 사용되는 특정 피처를 강조하여 성능을 증강시킬 수 있다. 더욱이, ASR에 의해 사용되지 않는 스펙트럼 성분을 필터링하는 것은 분리된 신호의 신호 대 잡음비(SNR)를 개선할 수 있고, 이는 결국 ASR의 성능을 개선시킬 수 있다. 전형적인 범위는 4-8kHz이다. 이 파라미터의 최적화는 골든 섹션 검색과 같은 다양한 최적화 방법을 사용하여 수행할 수 있다. 이 파라미터는 연속이다. 골든 섹션 검색을 적용할 때, 알고리즘에 대한 입력은 오류 함수(519)와 차단 주파수 섹션의 초기 추측 예를 들어, 1000Hz 및 0.5 X 샘플링 속도이다. 출력은 최적의 필터링 파라미터이다.

- [0162] 각각의 마이크로폰에 대한 가중치 인자. 각각의 마이크로폰의 가중치 인자는 디믹싱과 관련이 있다. 이론적으로 특정 어레이에 있는 상이한 마이크의 감도는 최대 3dB로 유사해야 한다. 그러나 실제로는 상이한 마이크로폰의 감도 스펜(span)이 더 클 수 있다. 또한, 마이크의 감도는 먼지와 습기로 인해 시간 경과에 따라 변할 수 있다. 전형적인 범위는 0-10dB이다. 이것은 연속 파라미터이다. 이 파라미터의 최적화는 각각의 마이크로폰에 대한 가중치 인자가 있거나 없는 Nelder-mead와 같은 다양한 최적화 방법을 사용하여 수행할 수 있다. Nelder-mead 방법을 적용할 때 알고리즘에 대한 입력은 오류 함수(519)와 심플렉스 꼭짓점의 초기 추측이다. 예를 들어, 각각의 n-튜플의 크기는 마이크로폰 수이다 - N: (1,0,...,0,0), (0,0,...,0,1) 및 (1/N,1/N,...,1/N). 출력은 마이크로폰당 최적의 가중치이다.
- [0163] 마이크로폰의 수
- [0164] 마이크로폰의 수는 디믹싱과 관련이 있다. 마이크로폰의 수는 한편으로는 분리할 수 있는 소스의 수에 영향을 미치고 다른 한편으로는 복잡성과 수치 정밀도에 영향을 미친다. 실제 실험에서도 너무 많은 마이크로폰은 출력 SNR을 감소시킬 수 있음을 보여주었다. 전형적인 범위는 4-8이다. 이것은 이산 파라미터이다. 이 파라미터의 최적화는 각각의 마이크로폰에 대한 가중치 인자를 사용하여 그리드 검색 또는 Nelder-mead와 같은 다양한 최적화 방법을 사용하여 수행될 수 있다. 그리드 검색을 적용할 때 알고리즘에 대한 입력은 오류 함수(519)와 검색이 수행되는 마이크로폰의 수이다. 예를 들어, 4,5,6,7,8 마이크로폰. 출력은 최적의 마이크로폰의 수이다.
- [0165] 도 9는 방법(600)을 예시한다.
- [0166] 방법(600)은 소스 선택 프로세스의 이전 출력에 적용된 음성 인식 프로세스와 관련된 오류를 수신하거나 계산하는 단계(610)에 의해 시작할 수 있다.
- [0167] 단계 (610) 다음에는 오류에 기초하여 소스 분리 프로세스의 적어도 하나의 파라미터를 수정하는 단계 (620)가 뒤따를 수 있다.
- [0168] 단계 (620) 다음에는 다수의 소스들로부터 발원되고 마이크로폰들의 어레이에 의해 검출되는 오디오 신호들을 나타내는 신호들을 수신하는 단계 (630)가 뒤따를 수 있다.
- [0169] 단계 (630) 다음에는 소스 분리된 신호들을 제공하기 위해 다수의 소스들의 상이한 소스들로부터 발원된 오디오 신호들을 분리하기 위한 소스 분리 프로세스를 수행하는 단계; 및 소스 분리된 신호를 소스 선택 프로세스로 발송하는 단계(640)가 뒤따를 수 있다.
- [0170] 단계 (640) 다음에는 단계 (630)가 뒤따를 수 있다.
- [0171] 단계 (630 및 640)의 각각의 한번 이상의 반복 다음에는 단계 (610)가 뒤따를 수 있다(도시되지 않음) - 여기서 단계 (640)의 출력은 ASR의 이전 출력을 제공하기 위해 소스 선택 프로세스 및 ASR에 공급될 수 있다.
- [0172] 단계 (630 및 640)의 초기 반복은 오류를 수신하지 않고 실행될 수 있다는 것에 유의해야 한다.
- [0173] 단계 (640)은 주파수 변환 (FFT와 같으나 이에 제한되지 않음)을 적용하는 단계, 디믹싱하는 단계, 및 역 주파수 변환(IFTT와 같으나 이에 제한되지 않음)을 적용하는 단계를 포함할 수 있다.
- [0174] 단계 (620)은 다음 중 적어도 하나를 포함할 수 있다:
- [0175] a. 주파수 변환의 하나 이상의 파라미터를 수정하는 단계.
- [0176] b. 역 주파수 변환의 하나 이상의 파라미터를 수정하는 단계.
- [0177] c. 디믹싱의 하나 이상의 파라미터를 수정하는 단계.
- [0178] d. 주파수 변환이 적용되는 오디오 신호를 나타내는 신호 세그먼트의 길이를 수정하는 단계.
- [0179] e. 오디오 신호를 나타내는 신호의 연속 세그먼트 간의 중첩을 수정하는 단계, 주파수 변환은 세그먼트 기반으로 적용된다.
- [0180] f. 주파수 변환의 샘플링 속도를 수정하는 단계.
- [0181] g. 주파수 변환에 의해 적용된 윈도우의 윈도우 파라미터를 수정하는 단계.
- [0182] h. 디믹싱 동안에 적용되는 필터의 차단 주파수를 수정하는 단계.
- [0183] i. 디믹싱 동안에 마이크로폰 어레이에서 각각의 마이크로폰에 적용된 가중치를 수정하는 단계.

- [0184] j. 마이크로폰 어레이의 마이크로폰 수를 수정하는 단계.
- [0185] k. 골든 섹션 검색을 사용하여 적어도 하나의 파라미터의 수정된 값을 결정하는 단계.
- [0186] l. Nedler Mead 알고리즘을 사용하여 적어도 하나의 파라미터의 수정된 값을 결정하는 단계.
- [0187] m. 그리드 검색을 사용하여 적어도 하나의 파라미터의 수정된 값을 결정하는 단계.
- [0188] n. 상기 오류와 상기 적어도 하나의 파라미터 사이의 미리 정의된 매핑에 기초하여 적어도 하나의 파라미터의 파라미터의 수정된 값을 결정하는 단계.
- [0189] o. 오류와 하나 이상의 파라미터 간의 매핑을 실시간으로 결정하는 단계.
- [0190] 전술한 명세서에서, 본 발명은 본 발명의 구체예의 특정 예를 참조하여 설명되었다. 그러나, 첨부된 청구 범위 에 기재된 본 발명의 보다 넓은 사상 및 범위를 벗어나지 않고 다양한 수정 및 변경이 이루어질 수 있음이 명백 할 것이다.
- [0191] 또한, 설명 및 청구 범위 (있는 경우)에서 용어 "앞", "뒤", "상부", "바닥", "위", "아래" 등은 설명 목적으로 사용되며 영구적인 상대적 위치를 설명하기 위한 것은 아니다. 그렇게 사용된 용어는 본 출원에서 설명된 본 발 명의 실시예가, 예를 들어, 본 출원에 예시되거나 달리 기술된 것과 다른 방위에서 동작할 수 있도록 적절한 상 황에서 상호 교환가능 하다는 것이 이해된다.
- [0192] 동일한 기능을 달성하기 위한 컴포넌트의 임의의 배열은 원하는 기능이 달성되도록 효과적으로 "연관"된다. 따 라서, 특정 기능을 달성하기 위해 조합된 본 출원의 임의의 2개의 컴포넌트는 아키텍처 또는 중간 컴포넌트와 상관없이 원하는 기능이 달성되도록 서로 "연관된" 것으로 보일 수 있다. 마찬가지로, 그렇게 연관된 임의의 두 컴포넌트는 원하는 기능을 달성하기 위해 서로 "동작 가능하게 연결"되거나 "동작 가능하게 결합"된 것으로 볼 수 있다.
- [0193] 더욱이, 당업자는 상기에서 설명된 동작들 사이의 경계가 단지 예시적이라는 것을 인식할 것이다. 다수의 동작 은 단일 동작으로 조합될 수 있고, 단일 동작은 추가 동작으로 분산될 수 있으며 동작은 시간상 적어도 부분적 으로 중첩되어 실행될 수 있다. 더욱이, 대안적인 실시예는 특정 동작의 다수의 인스턴스를 포함할 수 있고, 동 작의 순서는 다양한 다른 실시예에서 변경될 수 있다.
- [0194] 그러나, 다른 수정, 변형 및 대안도 가능하다. 따라서 명세서 및 도면은 제한적인 의미가 아니라 예시적인 것으 로 간주되어야 한다.
- [0195] "X일 수 있음"이라는 문구는 조건 X가 충족될 수 있음을 나타낸다. 이 어구는 또한 조건 X가 충족되지 않을 수 있음을 암시한다. 예를 들어, 특정 컴포넌트를 포함하는 시스템에 대한 임의의 언급은 시스템에 특정 컴포넌트 가 포함되지 않은 시나리오도 커버해야 한다. 예를 들어, - 특정 단계를 포함하는 방법에 대한 임의의 언급은 방법에 특정 컴포넌트가 포함되지 않은 시나리오도 커버해야 한다. 또 다른 예로, 특정 작업을 수행하도록 구성 된 시스템에 대한 임의의 언급은 시스템이 특정 동작을 수행하도록 구성되지 않은 시나리오도 커버해야 한다.
- [0196] "포함하는", "포함하는", "갖는", "구성되는" 및 "본질적으로 구성되는"이라는 용어는 상호 교환 가능한 방식으 로 사용된다. 예를 들어, 임의의 방법은 적어도 도면 및/또는 명세서에 포함된 단계, 도면 및/또는 명세서에 포 함된 단계만 포함할 수 있다. 시스템에도 동일하게 적용된다.
- [0197] 시스템은 마이크로폰 어레이, 메모리 유닛 및 디지털 신호 프로세서, FPGA, ASIC, 위에서 언급한 방법 중 임의 의 것을 실행하도록 프로그래밍된 범용 프로세서와 같은 하나 이상의 하드웨어 프로세서 등을 포함할 수 있다. 시스템은 마이크로폰 어레이를 포함하지 않을 수 있지만 마이크로폰 어레이에서 생성된 사운드 신호로부터 공급 될 수 있다.
- [0198] 예시의 단순성과 명료성을 위해, 도면에 도시된 엘리먼트는 반드시 축척에 맞게 그려지지 않는다는 것을 이해 할 것이다. 예를 들어, 일부 엘리먼트의 치수는 명확성을 위해 다른 엘리먼트에 비해 과장될 수 있다. 또한, 적 절하다고 판단되는 경우, 대응하는 또는 유사한 엘리먼트를 나타내기 위해 도면 사이에 참조 번호가 반복될 수 있다.
- [0199] 전술한 명세서에서, 본 발명은 본 발명의 구체예의 특정 예를 참조하여 설명되었다. 그러나, 첨부된 청구 범위 에 기재된 본 발명의 보다 넓은 사상 및 범위를 벗어나지 않고 다양한 수정 및 변경이 이루어질 수 있음이 명백 할 것이다.

- [0200] 또한, 설명 및 청구 범위(있는 경우)에서 용어 "앞", "뒤", "상부", "바닥", "위", "아래" 등은 설명 목적으로 사용되며 영구적인 상대적 위치를 설명하기 위한 것은 아니다. 그렇게 사용된 용어는 본 출원에 기술된 본 발명의 실시예가, 예를 들어, 본 출원에 예시되거나 달리 기술된 것과 다른 방위에서 동작할 수 있도록 적절한 상황에서 상호 교환 가능하다는 것이 이해된다.
- [0201] 당업자는 로직 블록 사이의 경계가 단지 예시적이며 대안적인 실시예가 로직 블록 또는 회로 엘리먼트를 병합하거나 다양한 로직 블록 또는 회로 엘리먼트에 기능의 대안적인 분해를 부과할 수 있음을 인식할 것이다. 따라서, 본 출원에 묘사된 아키텍처는 단지 예시적이며, 실제로 동일한 기능을 달성하는 많은 다른 아키텍처가 구현될 수 있다는 것을 이해해야 한다.
- [0202] 동일한 기능을 달성하기 위한 컴포넌트의 임의의 배열은 원하는 기능이 달성되도록 효과적으로 "연관"된다. 따라서, 특정 기능을 달성하기 위해 본 출원에 결합된 임의의 두 컴포넌트는 아키텍처 또는 중간 컴포넌트에 관계 없이 원하는 기능이 달성되도록 서로 "연관된" 것으로 볼 수 있다. 유사하게, 그렇게 연관된 임의의 2개의 컴포넌트는 또한 원하는 기능을 달성하기 위해 서로 "동작 가능하게 연결"되거나 "동작 가능하게 결합"된 것으로 볼 수 있다.
- [0203] 더욱이, 당업자는 상기에서 설명된 동작들 사이의 경계가 단지 예시적이라는 것을 인식할 것이다. 다수의 동작은 단일 동작으로 조합될 수 있고, 단일 동작은 추가 동작으로 분산될 수 있으며 동작은 시간상 적어도 부분적으로 중첩되어 실행될 수 있다. 더욱이, 대안적인 실시예는 특정 동작의 다중 인스턴스를 포함할 수 있고, 동작의 순서는 다양한 다른 실시예에서 변경될 수 있다.
- [0204] 또한 예를 들어, 일 실시예에서, 예시된 예는 단일 집적 회로 상에 또는 동일한 디바이스 내에 위치한 회로부로서 구현될 수 있다. 대안적으로, 예는 적절한 방식으로 서로 상호 연결된 임의의 수의 개별 집적 회로 또는 개별 디바이스로서 구현될 수 있다.
- [0205] 또한 예를 들어, 예 또는 그 일부는 임의의 적절한 유형의 하드웨어 설명 언어에서와 같이 물리적 회로부 또는 물리적 회로로 변환 가능한 논리적 표현의 소프트 또는 코드 표현으로 구현될 수 있다.
- [0206] 또한, 본 발명은 프로그래밍할 수 없는 하드웨어로 구현된 물리적 디바이스 또는 유닛에 제한되지 않고 본 출원에서 '컴퓨터 시스템'으로 일반적으로 표시된 메인 프레임, 미니컴퓨터, 서버, 워크스테이션, 개인용 컴퓨터, 노트북, 개인 휴대 정보 단말기, 전자 게임, 자동차 및 기타 임베디드 시스템, 휴대폰 및 다른 다양한 무선 디바이스와 같은 적절한 프로그램 코드에 따라 동작함으로써 원하는 디바이스 기능을 수행할 수 있는 프로그래밍 가능한 디바이스 또는 유닛에도 적용될 수 있다.
- [0207] 그러나, 다른 수정, 변형 및 대안도 가능하다. 따라서 명세서 및 도면은 제한적인 의미가 아니라 예시적인 것으로 간주되어야 한다.
- [0208] 청구 범위에서, 괄호 사이에 배치된 임의의 참조 부호는 청구 범위를 제한하는 것으로 해석되어서는 안 된다. '포함하는'이라는 단어는 청구 범위에 나열된 것 외에 다른 엘리먼트 또는 단계의 존재를 배제하지 않는다. 또한, 본 출원에 사용된 용어 "a" 또는 "an"은 하나 또는 하나 이상으로 정의된다. 또한, 청구 범위에서 "적어도 하나" 및 "하나 이상"과 같은 도입구의 사용은 부정관사 "a" 또는 "an"에 의한 다른 청구 엘리먼트의 도입이 동일한 청구항이 "하나 이상" 또는 "적어도 하나"의 도입구와 "a" 또는 "an"과 같은 부정관사를 포함하는 경우에도 이러한 도입된 청구항 엘리먼트를 포함하는 임의의 특정 청구항을 그러한 엘리먼트를 하나만 포함하는 발명으로 제한한다는 의미로 해석되어서는 안 된다. 정관사를 사용하는 경우에 마찬가지이다. 달리 명시되지 않는 한, "제1" 및 "제2"와 같은 용어는 해당 용어가 설명하는 엘리먼트를 임의로 구별하는 데 사용된다. 따라서 이러한 용어는 특정 조치가 상이한 청구 범위에 인용되어 있다는 사실이 이러한 조치의 조합이 유리하게 사용될 수 없다는 것을 나타내지 않는다는 단순한 사실만으로 그러한 엘리먼트의 시간적 또는 다른 우선순위 지정을 나타내도록 의도된 것은 아니다.
- [0209] 본 발명은 또한 컴퓨터 시스템과 같은 프로그램 가능한 장치에서 실행될 때 본 발명에 따른 방법의 단계를 수행하거나 또는 프로그램 가능한 장치가 본 발명에 따른 디바이스 또는 시스템의 기능을 수행하는 것을 가능하게 하기 위한 코드 부분을 적어도 포함하는 컴퓨터 시스템에서 실행하기 위한 컴퓨터 프로그램으로 구현될 수 있다. 컴퓨터 프로그램으로 인해 스토리지 시스템이 디스크 드라이브를 디스크 드라이브 그룹에 할당할 수 있다.
- [0210] 컴퓨터 프로그램은 특정 애플리케이션 및/또는 운영 체제와 같은 명령 리스트이다. 컴퓨터 프로그램은 예를 들

어, 서브루틴, 함수, 프로시저(procedure), 객체 방법, 객체 구현, 실행 가능한 애플리케이션, 애플릿, 서블릿(servlet), 소스 코드, 객체 코드, 공유 라이브러리/동적 로드 라이브러리 및/또는 컴퓨터 시스템에서 실행되도록 디자인된 기타 명령 시퀀스 중 하나 이상을 포함할 수 있다.

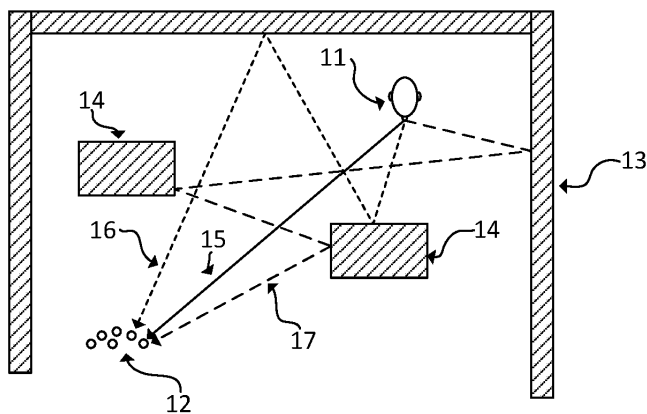
[0211] 컴퓨터 프로그램은 비일시적 컴퓨터 판독 가능 매체에 내부적으로 저장될 수 있다. 컴퓨터 프로그램의 전부 또는 일부는 정보 처리 시스템에 영구적으로, 제거 가능하게 또는 원격으로 결합된 컴퓨터 판독 가능 매체에 제공될 수 있다. 컴퓨터 판독 가능 매체는 예를 들어, 제한 없이 다음 중 임의의 개수를 포함할 수 있다: 디스크 및 테이프 저장 매체를 포함하는 자기 스토리지; 광 저장 매체 예컨대, 콤팩트 디스크 미디어(예를 들어, CD ROM, CD R 등) 및 디지털 비디오 디스크 저장 매체; FLASH 메모리, EEPROM, EPROM, ROM과 같은 반도체 기반 메모리 유닛을 포함하는 휘발성 메모리 저장 매체; 강자성 디지털 메모리; MRAM; 레지스터, 버퍼 또는 캐시, 메인 메모리, RAM 등을 포함하는 휘발성 저장 매체. 컴퓨터 프로세스는 전형적으로 프로세스의 실행을 관리하기 위해 수행(실행 중인) 프로그램 또는 프로그램의 일부, 현재 프로그램 값 및 상태 정보, 운영 체제에서 사용하는 자원을 포함한다. 운영 체제(OS)는 컴퓨터 자원 공유를 관리하고 프로그래머에게 해당 자원에 액세스하는 데 사용되는 인터페이스를 제공하는 소프트웨어이다. 운영체제는 시스템 데이터와 사용자 입력을 처리하고, 시스템의 사용자와 프로그램에 대한 서비스로서 작업과 내부 시스템 자원을 할당 및 관리함으로써 응답한다. 컴퓨터 시스템은 예를 들어, 적어도 하나의 처리 유닛, 관련 메모리 및 다수의 입출력(I/O) 디바이스를 포함할 수 있다. 컴퓨터 프로그램을 실행할 때, 컴퓨터 시스템은 컴퓨터 프로그램에 따라 정보를 처리하고 I/O 디바이스를 통해 결과 출력 정보를 생성한다.

[0212] 이 특허 출원에 언급된 임의의 시스템은 적어도 하나의 하드웨어 컴포넌트를 포함한다.

[0213] 본 발명의 특정 피처가 본 출원에 예시되고 설명되었지만, 이제 많은 수정, 대체, 변경 및 균등물이 당업자에게 일어날 것이다. 따라서, 첨부된 청구 범위는 본 발명의 진정한 사상에 속하는 모든 그러한 수정 및 변경을 포함하도록 의도된 것으로 이해되어야 한다.

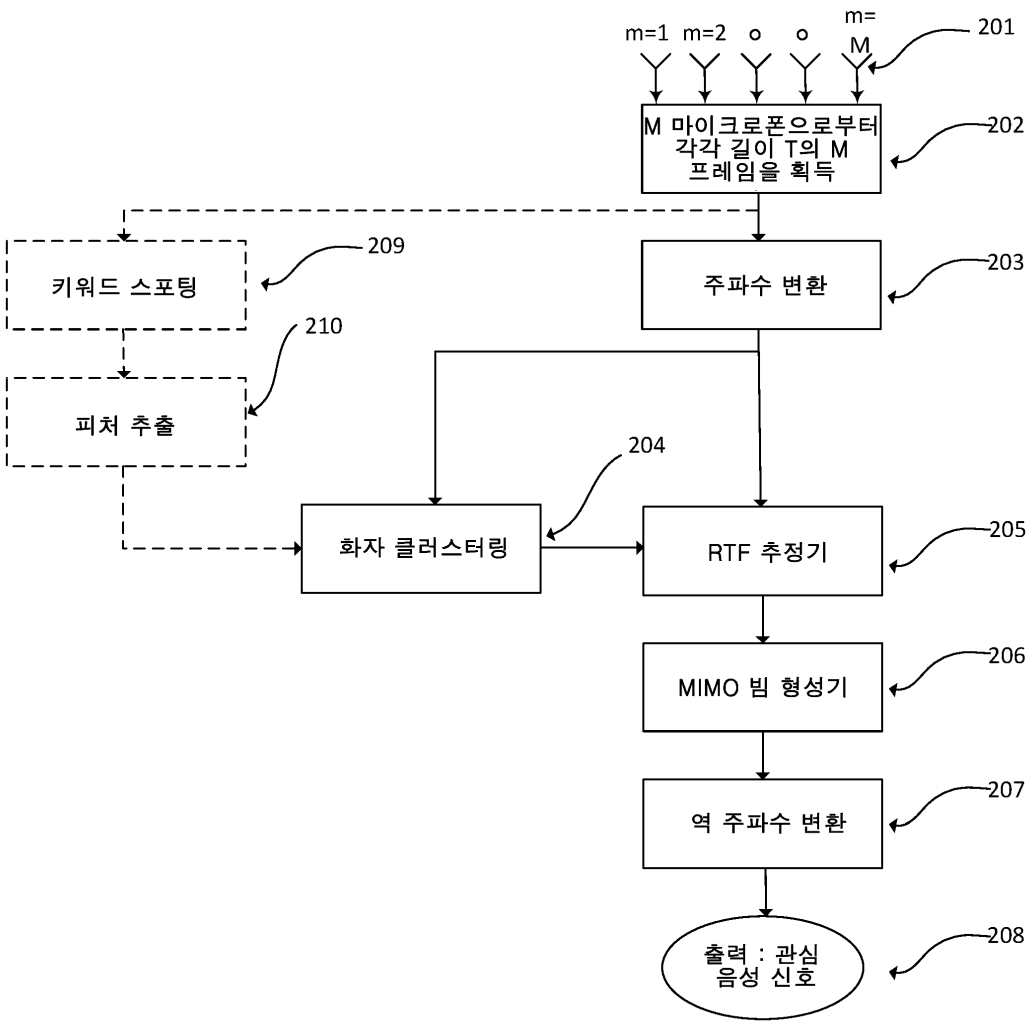
## 도면

### 도면1



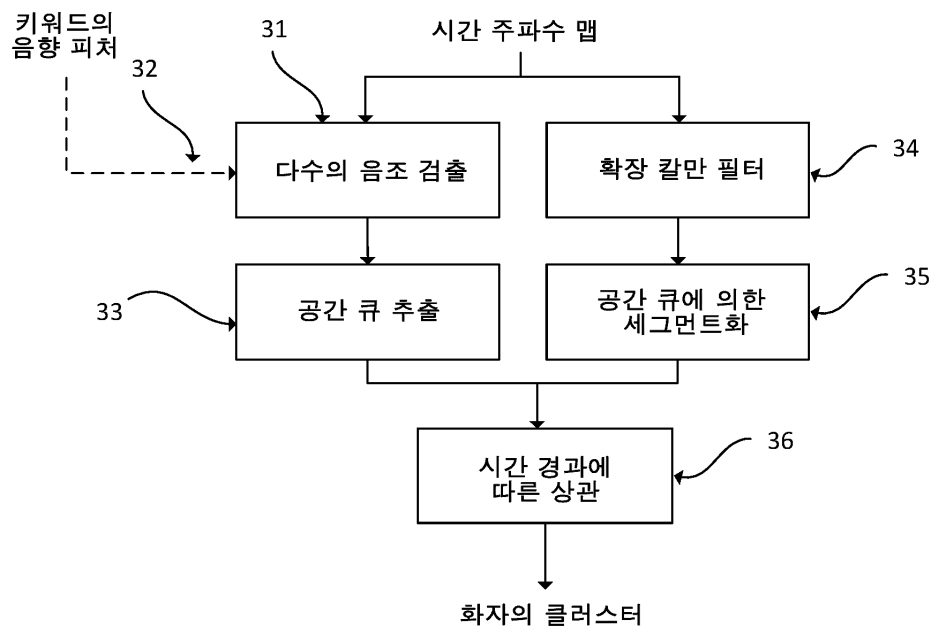
10

도면2



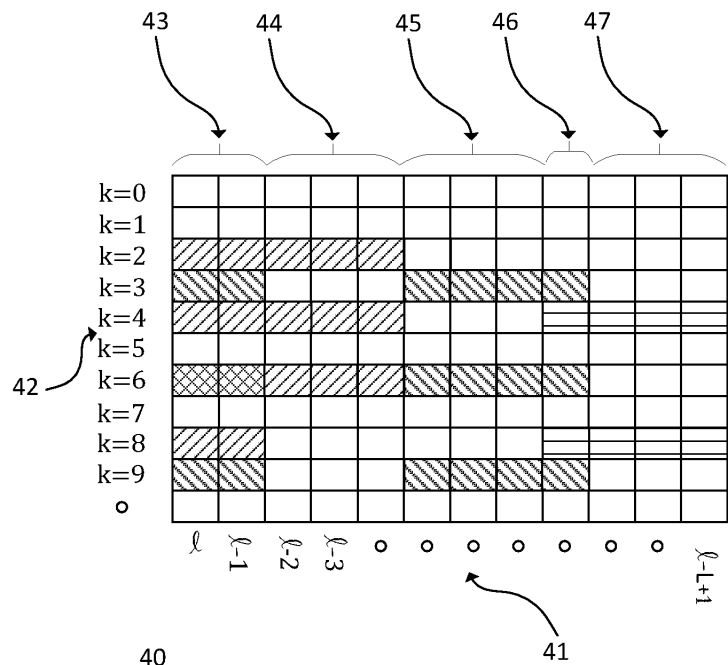
200

도면3



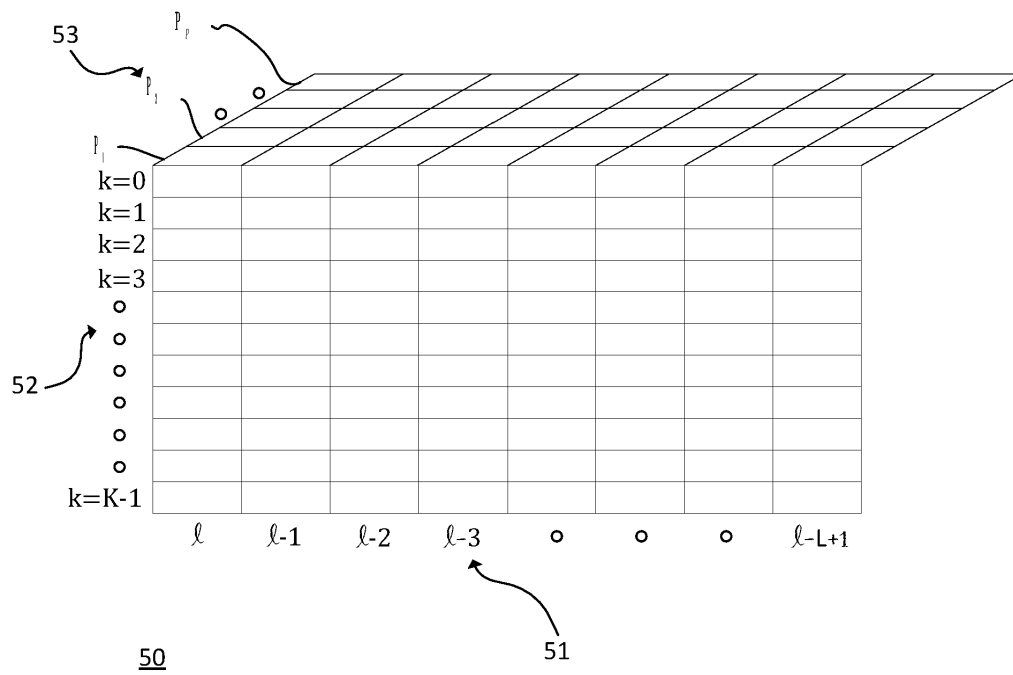
204

도면4

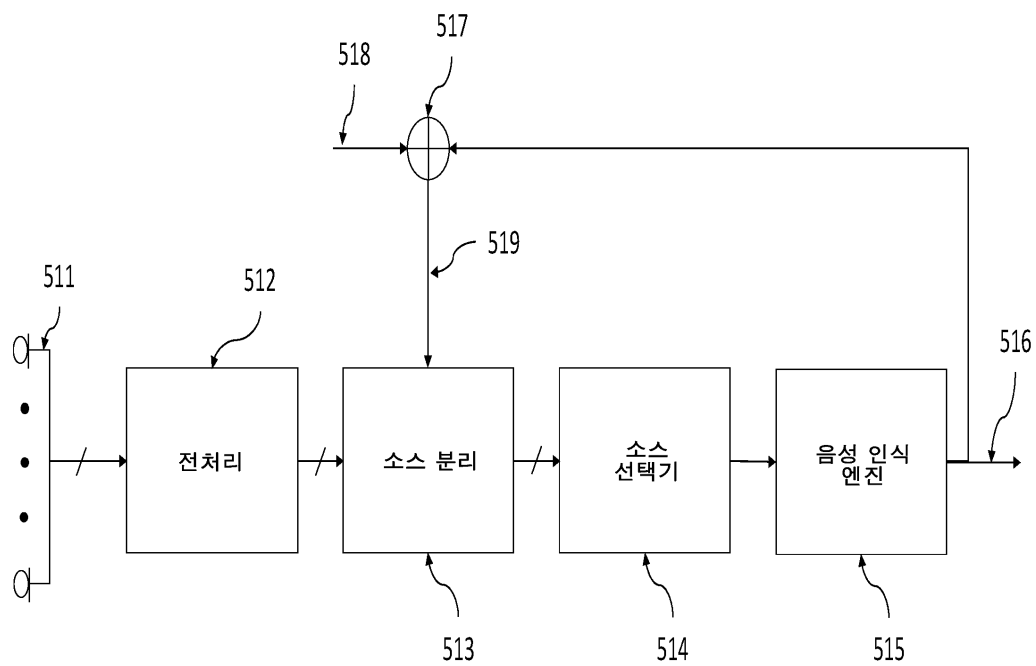


40

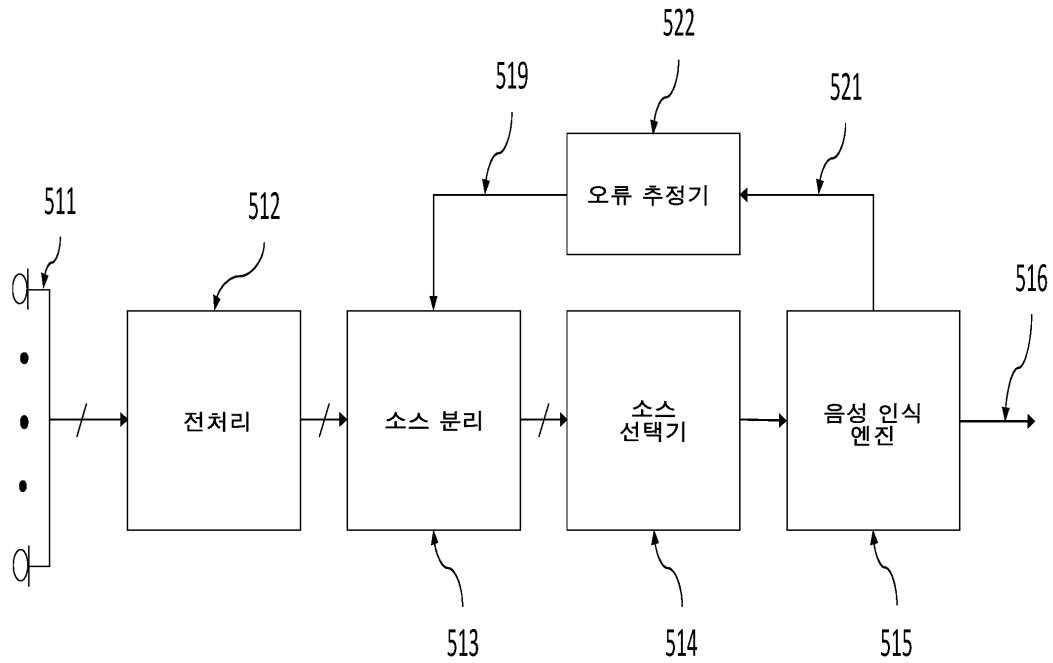
도면5



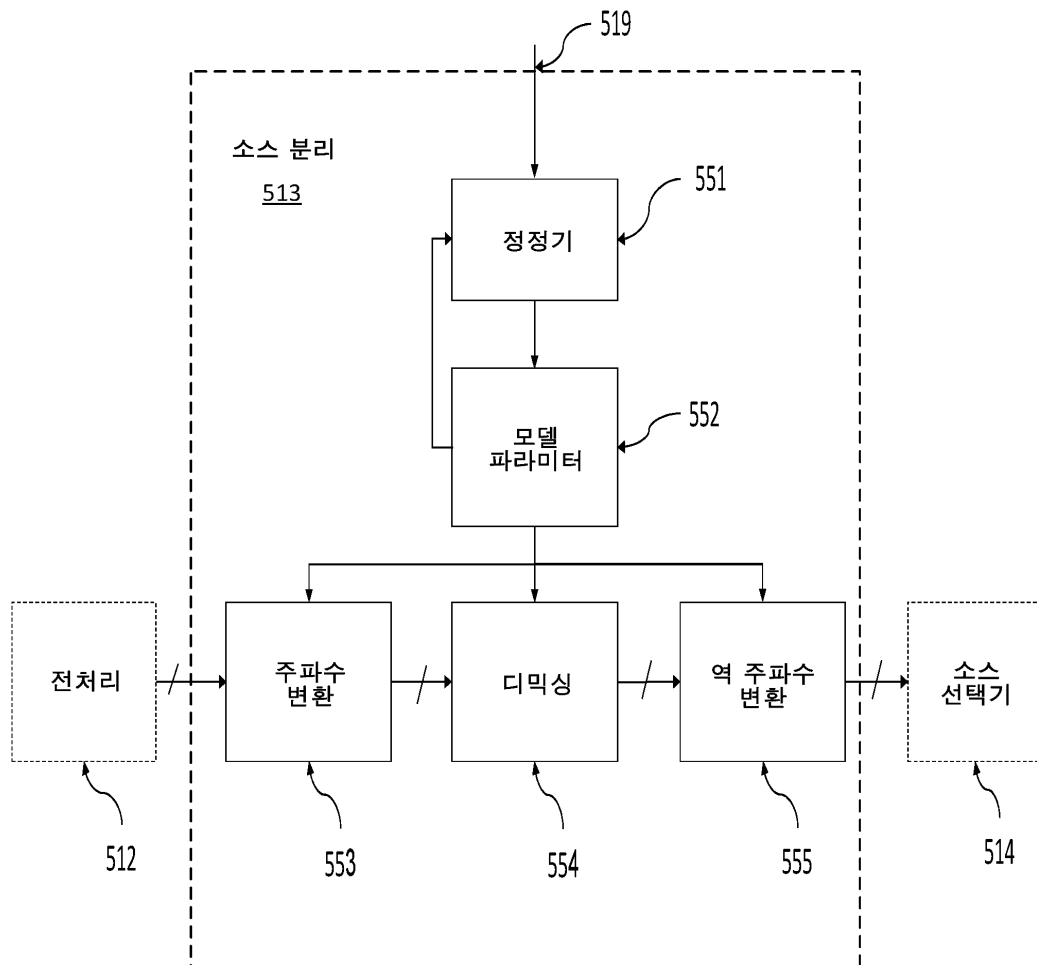
도면6



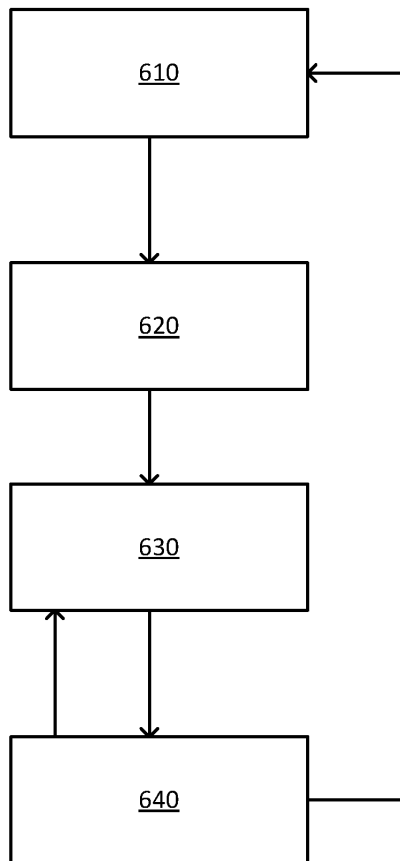
도면7



도면8



도면9



600