(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2012-0102306
(43) 공개일자 2012년09월18일

(51) 국제특허분류(Int. Cl.)
G10L 21/02 (2006.01) *G10L 15/00* (2006.01)
(21) 출원번호 10-2011-0020385
(22) 출원일자 2011년03월08일
심사청구일자 2011년03월08일

(71) 출원인
고려대학교 산학협력단
서울 성북구 안암동5가 1
(72) 발명자
임현택
경기도 안산시 상록구 예술광장1로 131, 선경아파트 17동 801호 (성포동)
육동석
서울특별시 용산구 이촌로 347, 1동 102호 (서빙고동, 신동아아파트)
(74) 대리인
특허법인충현

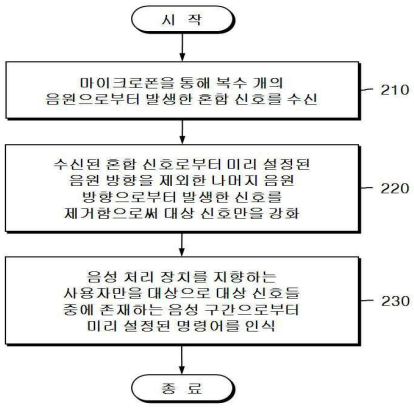
전체 청구항 수 : 총 19 항

(54) 발명의 명칭 **잡음 환경에서의 음성 처리 장치 및 방법**

(57) 요 약

본 발명은 잡음 환경에서의 음성 처리 장치, 방법 및 그 방법을 기록한 기록매체에 관한 것으로, 본 발명에 따른 음성 처리 방법은 마이크로폰을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신하고, 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화하며, 음성 처리 장치를 지향하는 사용자만을 대상으로 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식한다.

대 표 도 - 도2



특허청구의 범위

청구항 1

마이크로폰을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신하는 단계;

상기 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화하는 단계; 및

음성 처리 장치를 지향하는 사용자만을 대상으로 상기 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식하는 단계를 포함하는 음성 처리 방법.

청구항 2

제 1 항에 있어서,

상기 명령어를 인식하는 단계는,

상기 강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출하는 단계;

상기 검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정하는 단계; 및

상기 결정된 음성 인식 대상의 음성 구간으로부터 상기 미리 설정된 명령어를 탐지하는 단계를 포함하는 음성 처리 방법.

청구항 3

제 2 항에 있어서,

상기 음성 구간을 검출하는 단계는,

모음의 스펙트럼에서 특징 피크가 위치한 피크 대역을 나타내는 모음 특징 정보를 미리 저장하는 단계; 및

상기 강화된 대상 신호들의 스펙트럼에서 상기 저장된 모음 특징 정보가 나타내는 피크 대역에 대응하는 대응 대역 및 상기 대응 대역을 제외한 비대응 대역의 평균 에너지를 이용하여 상기 강화된 대상 신호들 중 음성에 해당하는 구간을 검출하는 단계를 포함하는 음성 처리 방법.

청구항 4

제 2 항에 있어서,

상기 사용자를 음성 인식 대상으로 선택적으로 결정하는 단계는,

안면 인식을 이용하여 상기 사용자가 상기 음성 처리 장치를 지향하고 있는지를 판단하는 단계;

상기 안면 인식 결과에 기초하여 상기 사용자가 미리 등록된 사용자인지 여부를 검사하는 단계; 및

상기 판단 결과 및 상기 검사 결과를 모두 만족하는 사용자를 상기 음성 인식 대상으로 결정하는 단계를 포함하는 음성 처리 방법.

청구항 5

제 2 항에 있어서,

상기 사용자를 음성 인식 대상으로 선택적으로 결정하는 단계는,

상기 사용자의 음성 신호 및 은닉 마코브 모델(hidden Markov models, HMM)을 이용한 화자 인식 결과에 기초하여 상기 사용자가 미리 등록된 사용자인지 여부를 검사하는 단계; 및

상기 검사 결과를 만족하는 사용자를 상기 음성 인식 대상으로 결정하는 단계를 포함하는 음성 처리 방법.

청구항 6

제 2 항에 있어서,

상기 미리 설정된 명령어를 탐지하는 단계는,

사용자의 의해 발음될 수 있는 음절의 대표 발음들로 구성되는 음절 모델과 명령어 세트를 미리 저장하는 단계;

상기 결정된 음성 인식 대상의 음성 구간으로부터 각 음절의 특징을 추출하는 단계;

상기 추출된 특징을 상기 저장된 음절 모델과 비교하여 적어도 하나 이상의 음절 후보를 생성하는 단계; 및

상기 생성된 음절 후보로부터 조합 가능한 어절들 중 상기 저장된 명령어 세트에 포함되는 명령어를 탐지하는 단계를 포함하는 음성 처리 방법.

청구항 7

제 1 항에 있어서,

상기 대상 신호만을 강화하는 단계는,

상기 음성 처리 장치로부터 발생되어 상기 마이크로폰을 통해 입력되는 에코(echo) 신호를 제거하는 단계;

상기 수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색하는 단계; 및

상기 탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거하는 단계를 포함하는 음성 처리 방법.

청구항 8

제 7 항에 있어서,

상기 신호들의 방향을 탐색하는 단계는, 도착 시간 지연법(TDOA, time delay of arrival), 빔 형성 방법(beam-forming), 고해상도 스펙트럼 추정 방법(spectral analysis) 중 적어도 하나를 이용하여 수행되는 것을 특징으로 하는 음성 처리 방법.

청구항 9

제 7 항에 있어서,

상기 잡음 신호를 제거하는 단계는,

상기 신호들에 대한 빔 패턴(beam pattern) 중에서 상기 미리 설정된 음원 방향에 해당하는 메인 로브(main lobe)를 제외한 사이드 로브(side lobe)에 대한 제거 신호를 생성하는 단계; 및

상기 혼합 신호와 상기 생성된 제거 신호를 합성하는 단계를 포함하는 음성 처리 방법.

청구항 10

제 1 항 내지 제 9 항 중에 어느 한 항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 11

마이크로폰을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신하는 수신부;

상기 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화하는 신호 강화부; 및

음성 처리 장치를 지향하는 사용자만을 대상으로 상기 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식하는 음성 명령어 인식부를 포함하는 음성 처리 장치.

청구항 12

제 11 항에 있어서,

상기 음성 명령어 인식부는,

상기 강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출하는 음성 검출부;

상기 검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정하는 대상 결정부; 및

상기 결정된 음성 인식 대상의 음성 구간으로부터 상기 미리 설정된 명령어를 탐지하는 명령어 탐지부를 포함하는 음성 처리 장치.

청구항 13

제 12 항에 있어서,

상기 음성 검출부는,

모음의 스펙트럼에서 특징 피크가 위치한 피크 대역을 나타내는 모음 특징 정보를 미리 저장하고,

상기 강화된 대상 신호들의 스펙트럼에서 상기 저장된 모음 특징 정보가 나타내는 피크 대역에 대응하는 대응 대역 및 상기 대응 대역을 제외한 비대응 대역의 평균 에너지를 이용하여 상기 강화된 대상 신호들 중 음성에 해당하는 구간을 검출하는 것을 특징으로 하는 음성 처리 장치.

청구항 14

제 12 항에 있어서,

상기 대상 결정부는,

안면 인식을 이용하여 상기 사용자가 상기 음성 처리 장치를 지향하고 있는지를 판단하고,

상기 안면 인식 결과에 기초하여 상기 사용자가 미리 등록된 사용자인지 여부를 검사하며,

상기 판단 결과 및 상기 검사 결과를 모두 만족하는 사용자를 상기 음성 인식 대상으로 결정하는 것을 특징으로 하는 음성 처리 장치.

청구항 15

제 12 항에 있어서,

상기 대상 결정부는,

상기 사용자의 음성 신호 및 은닉 마코브 모델을 이용한 화자 인식 결과에 기초하여 상기 사용자가 미리 등록된 사용자인지 여부를 검사하고,

상기 검사 결과를 만족하는 사용자를 상기 음성 인식 대상으로 결정하는 것을 특징으로 하는 음성 처리 장치.

청구항 16

제 12 항에 있어서,

상기 명령어 탐지부는,

사용자의 의해 발음될 수 있는 음절의 대표 발음들로 구성되는 음절 모델과 명령어 세트를 미리 저장하고,

상기 결정된 음성 인식 대상의 음성 구간으로부터 각 음절의 특징을 추출하고,

상기 추출된 특징을 상기 저장된 음절 모델과 비교하여 적어도 하나 이상의 음절 후보를 생성하며,

상기 생성된 음절 후보로부터 조합 가능한 어절들 중 상기 저장된 명령어 세트에 포함되는 명령어를 탐지하는 것을 특징으로 하는 음성 처리 장치.

청구항 17

제 11 항에 있어서,

상기 신호 강화부는,

상기 음성 처리 장치로부터 발생되어 상기 마이크로폰을 통해 입력되는 에코 신호를 제거하는 에코 제거부;

상기 수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색하는 방향 탐색부; 및

상기 탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거하는 잡음 제거부를 포함하는 음성 처리 장치.

청구항 18

제 17 항에 있어서,

상기 방향 탐색부는, 도착 시간 지연법, 빔 형성 방법, 고해상도 스펙트럼 추정 방법 중 적어도 하나를 이용하여 상기 신호들의 방향을 탐색하는 것을 특징으로 하는 음성 처리 장치.

청구항 19

제 17 항에 있어서,

상기 잡음 제거부는,

상기 신호들에 대한 빔 패턴 중에서 상기 미리 설정된 음원 방향에 해당하는 메인 로브를 제외한 사이드 로브에 대한 제거 신호를 생성하고,

상기 혼합 신호와 상기 생성된 제거 신호를 합성하는 것을 특징으로 하는 음성 처리 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 잡음 환경에서의 음성 처리 장치 및 방법에 관한 것으로, 특히 사용자의 음성 및 기타 잡음이 혼재되어 있는 환경 하에서 대상 기기가 사용자의 명령을 인식함으로써 사용자로 하여금 별도의 버튼 조작 없이 대상 기기를 조작, 제어하는 장치, 방법 및 그 방법을 기록한 기록매체에 관한 것이다.

배 경 기 술

[0002] 음성 인식은 음성 처리 기술의 발달과 함께 오래 전부터 연구되어 온 주제이다. 음성 인식과 관련하여 인식률을 향상시키기 위한 다양한 연구들이 진행되어 왔고, 현재까지도 계속 연구되고 있다. 음성인식은 잡음이 존재하지 않는 실험적인 환경에서는 상대적으로 인식률이 높다. 하지만 음성 인식 기술이 실제 적용될 수 있는 환경에서는 주변에 존재하는 다양한 잡음, 예측 불가능한 화자의 상태, 채널의 특성 등의 문제로 인해 음성 인식률이 급격히 떨어지는 문제점이 존재한다. 그로 인해 사용자의 수동 조작의 불편함을 덜어 주고자 연구되어 온 음성 인식 기술이 비록 기존의 수동 조작 인터페이스에 비해 보다 직관적이고 쉽게 사용할 수 있는 방법임에도 불구하고 널리 이용되지 못하고 있는 것이 현실이다.

[0003] 한편, 음성 입출력 인터페이스를 이용하여 음성 입력을 처리하기 위해서는 기본적으로 해당 입력 장치에 음성 입력의 시작과 끝을 알려주는 시작 버튼과 종료버튼을 사용하여 음성 녹음 및 인식을 처리한다. 이러한 수동 버튼 조작과 같은 번거로운 문제점이 있기 때문에 실제 일반 사용자가 자유롭게 음성 입력을 이용하는 것은 한계가 있다.

발명의 내용

해결하려는 과제

[0004] 본 발명이 해결하고자 하는 기술적 과제는 사용자의 음성과 다양한 잡음이 혼재되어 있는 환경 하에서 예측 불가능한 화자의 상태 및 채널의 특성으로 인해 음성 인식의 성능이 저하되는 문제점을 극복하고, 그 결과 다양한 기기들에서 음성 인식 기술이 활용되지 못하고 여전히 수동 조작에 의존하는 불편함을 해소하고자 한다. 한편, 본 발명이 해결하고자 하는 다른 기술적 과제는 음성 인식을 위한 음성 입력 과정에서 음성 입력의 시작과 끝을 수동 조작하는 번거로움을 제거하고자 한다.

과제의 해결 수단

[0005] 상기 기술적 과제를 해결하기 위하여, 본 발명에 따른 잡음 환경에서의 음성 처리 방법은 마이크로폰을 통해 복

수 개의 음원으로부터 발생한 혼합 신호를 수신하는 단계; 상기 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화하는 단계; 및 음성 처리 장치를 지향하는 사용자만을 대상으로 상기 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식하는 단계를 포함한다.

[0006] 또한, 상기된 잡음 환경에서의 음성 처리 방법에서 상기 대상 신호만을 강화하는 단계는, 상기 음성 처리 장치로부터 발생되어 상기 마이크로폰을 통해 입력되는 에코(echo) 신호를 제거하는 단계; 상기 수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색하는 단계; 및 상기 탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거하는 단계를 포함한다.

[0007] 나아가, 상기된 잡음 환경에서의 음성 처리 방법에서 상기 명령어를 인식하는 단계는, 상기 강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출하는 단계; 상기 검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정하는 단계; 및 상기 결정된 음성 인식 대상의 음성 구간으로부터 상기 미리 설정된 명령어를 탐지하는 단계를 포함한다.

[0008] 한편, 이하에서는 상기 기재된 잡음 환경에서의 음성 처리 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.

[0009] 상기 기술적 과제를 해결하기 위하여, 본 발명에 따른 잡음 환경에서의 음성 처리 장치는 마이크로폰을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신하는 수신부; 상기 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화하는 신호 강화부; 및 음성 처리 장치를 지향하는 사용자만을 대상으로 상기 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식하는 음성 명령어 인식부를 포함한다.

[0010] 또한, 상기된 잡음 환경에서의 음성 처리 장치에서 상기 신호 강화부는, 상기 음성 처리 장치로부터 발생되어 상기 마이크로폰을 통해 입력되는 에코 신호를 제거하는 에코 제거부; 상기 수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색하는 방향 탐색부; 및 상기 탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거하는 잡음 제거부를 포함한다.

[0011] 나아가, 상기된 잡음 환경에서의 음성 처리 장치에서 상기 음성 명령어 인식부는, 상기 강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출하는 음성 검출부; 상기 검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정하는 대상 결정부; 및 상기 결정된 음성 인식 대상의 음성 구간으로부터 상기 미리 설정된 명령어를 탐지하는 명령어 탐지부를 포함한다.

발명의 효과

[0012] 본 발명은 음성 신호와 잡음이 포함된 혼합 신호로부터 대상 음성 신호를 추출하여 명령어를 인식함으로써 음성 인식의 성능을 향상시키고, 수동 조작 없이 대상 기기를 용이하게 조작, 제어할 수 있다. 또한, 본 발명은 음성 인식을 위한 음성 입력 과정에서 모음 특징을 이용하여 음성 구간을 검출함으로써 별도의 번거로운 수동 조작 없이도 연속적인 음성 신호에서 명령어를 추출, 인식할 수 있다.

도면의 간단한 설명

[0013] 도 1은 잡음 환경에서 본 발명의 실시예들이 구현되는 예시 상황 및 기본 아이디어를 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 잡음 환경에서의 음성 처리 방법을 도시한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 도 2의 음성 처리 방법에서 대상 신호를 강화하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

도 4는 본 발명의 일 실시예에 따른 도 3의 음성 처리 방법에서 잡음 신호를 제거하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

도 5는 본 발명의 일 실시예에 따른 도 2의 음성 처리 방법에서 명령어를 인식하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

도 6은 본 발명의 일 실시예에 따른 도 5의 음성 처리 방법에서 음성 구간을 검출하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

도 7a 및 도 7b는 각각 본 발명의 일 실시예에 따른 도 5의 음성 처리 방법에서 사용자를 음성 인식 대상으로 선택적으로 결정하는 2가지 실시예들을 도시한 흐름도이다.

도 8은 본 발명의 일 실시예에 따른 도 5의 음성 처리 방법에서 미리 설정된 명령어를 탐지하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

도 9는 본 발명의 일 실시예에 따른 잡음 환경에서의 음성 처리 장치를 도시한 블록도이다.

도 10은 본 발명의 일 실시예에 따른 도 9의 음성 처리 장치에서 신호 강화부를 보다 구체적으로 도시한 블록도이다.

도 11은 본 발명의 일 실시예에 따른 도 9의 음성 처리 장치에서 음성 명령어 인식부를 보다 구체적으로 도시한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0014] 본 발명의 실시예들을 설명하기에 앞서 실시예들이 구현되는 환경에 대해 개괄적으로 소개하고, 실시예들이 공통적으로 채용하고 있는 기본 아이디어를 제시하고자 한다.

[0015] 도 1은 잡음 환경에서 본 발명의 실시예들이 구현되는 예시 상황 및 기본 아이디어를 설명하기 위한 도면으로서, TV를 중심으로 좌, 우에 다수의 사용자들이 배치되어 있으며, 사람들의 대화 소리, TV에서 발생하는 소리를 비롯하여 다양한 잡음이 혼재되어 있는 상황이라고 가정하자.

[0016] 본 발명의 다양한 실시예들은 일반적으로 음성 인식의 대상이 되는 대상 기기(도 1에서는 TV에 해당한다.)에 음성을 통해 명령을 내리게 되는 경우, 그 발화자는 대상 기기를 바라보는 사람일 것이며, 대상 기기로부터 일정 영역 내에 존재하는 사람일 것을 가정한다. 또한, 대상 기기에서 활용될 수 있는 명령어는 미리 결정되어 저장되어 있는 특정 명령어일 것을 전제로 한다. 따라서, 본 발명의 실시예들은 잡음 환경 하에서, 대상 기기 내의 특정 영역 내에 존재하고, 대상 기기를 바라보는 발화자의 음성 신호만을 그 음성 인식의 대상으로 한다.

[0017] 도 1에서, 도면 번호가 부여된 4명의 사용자(110, 120, 130, 140)에게 주목하자. 가장 좌측에 위치한 사용자(110)는 TV로부터 지나치게 측면에 위치해 있어 시청 각도가 적절하지 않으므로, 음성 인식의 대상에서 제외시킬 수 있을 것이다. 물론 대상 기기의 성격에 따라서는 음성 인식의 대상에 포함시킬 수도 있을 것이다. 또는, 화면 상에 보이지 않는 다른 사용자들의 경우에도 비록 발화하고 있는 사용자가 존재할지라도 그들은 음성 인식의 대상이 되지 않을 것이다. 두 번째 사용자(120)는 대상 기기인 TV를 바라보고 있으며, 유효한 시청 각도 내에 위치해 있으므로 음성 인식의 대상이 될 수 있을 것이다. 세 번째 사용자(130) 및 네 번째 사용자(140) 역시 대상 기기인 TV와의 유효한 시청 각도 내에는 존재하나, TV를 지향하고 있지 않으므로 음성 인식의 대상이 될 수 없을 것이다.

[0018] 이러한 기본 아이디어를 이용하여 다수의 음성과 잡음이 혼재되어 있는 환경 하에서 음성 인식의 대상이 되지 않는 신호들을 제거함으로써 보다 정확하고 빠르게 대상 기기를 목적으로 하는 음성 명령어를 인식할 수 있다. 즉, 본 발명의 실시예들은 음성 인식의 성능을 높이기 위해서 다양한 객체로부터 발생하는 음성 신호에 대한 왜곡을 보정하여 음성 인식기(음성 인식 대상 기기를 의미한다.)의 성능을 향상시키고자 한다. 또한, 음성 인터페이스를 사용하는데 있어서 가장 번거로운 문제인 음성의 시작과 끝을 사용자가 설정(Push to Talk)하는 과정 없이, 이러한 과정을 버튼과 같은 역할을 음성(키워드가 될 수 있다.)으로 대체함으로써 사용자의 편의성 및 음성 인터페이스의 사용을 증대시키고자 한다.

[0019] 이하에서, 관련 도면들을 참조하여 본 발명의 실시예들을 보다 구체적으로 설명한다. 도면들에서 동일한 참조 번호들은 동일한 구성 요소를 지칭한다.

[0020] 도 2는 본 발명의 일 실시예에 따른 잡음 환경에서의 음성 처리 방법을 도시한 흐름도로서, 다음과 같은 단계들을 포함한다.

[0021] 210 단계에서 마이크로폰(microphone)을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신한다. 마이크로폰은 혼합 신호들을 입력받는 장치로서, 통상의 단일 마이크로폰으로도 구현이 가능할 것이나, 각기 다른 다수의 음원들로부터 발생한 혼합 신호들을 수집하고, 수집된 신호를 용이하게 가공하기 위해서는 복수 개의 마이크로폰으로 구성되는 마이크로폰 어레이(array)인 것이 좀 더 유리할 것이다.

[0022] 220 단계에서 210 단계를 통해 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로

부터 발생한 신호를 제거함으로써 대상 신호만을 강화한다. 이러한 220 단계는 다수의 사용자 및 다양한 노이즈가 존재하는 환경에서 음성 인식의 성능을 높이기 위한 기초 신호를 생성하는 과정에 해당하는 것으로, 소리 신호를 입력받은 후 사람 이외의 음원으로부터 발생하는 소리(즉, 음성 신호를 제외한 기타 소리 신호를 의미한다.)를 포함하여 다양한 방향에서 존재하는 신호 중에서 원하는 신호를 찾아 음성 인식에 필요한 신호만을 강화하기 위한 일련의 과정들을 수행하게 된다. 각각의 과정들에 대해서는 이후에 도 3 및 도 4를 통해 구체적으로 설명하도록 한다.

[0023] 230 단계에서는 음성 처리 장치를 지향하는 사용자만을 대상으로 220 단계를 통해 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식한다. 이러한 230 단계는 기존의 음성 인터페이스 장치에 사용되었던 음성 처리의 시작과 끝을 알리기 위한 수동 조작을 자동화하기 위한 방법을 포함한다. 즉, 230 단계에서는 대상 기기(음성 인식을 하고자 하는 음성 처리 장치를 의미한다.)가 항상 사용자의 응답(음성 명령을 의미한다.)을 기다리면서 다양한 소리 및 다수의 사용자가 존재하는 환경에서 지정된 특정한 사용자의 명령을 지시하는 발화만을 음성 구간으로 판단하는 구조를 갖는다.

[0024] 이상과 같이 명령어 음성 구간으로부터 명령어를 인식하기 위해 대상 기기는 항시 사용자의 음성 명령에 대해 대기하며 오작동하지 않는 신뢰도를 보장할 수 있어야 한다. 이를 위해 본 발명의 실시예는 230 단계를 통해 음성 구간을 검출하고, 사용자의 안면을 인식하고, 발화자를 인식하며, 특정 키워드를 인식하는 일련의 과정을 수행한다. 이러한 일련의 과정은 구현 및 실시의 관점에서 성능 향상 또는 음성 인식 수행 속도의 향상을 위해 일부 단계의 가감이 가능할 것이다. 결과적으로 이상의 과정을 통해 음성 처리 장치는 입력된 소리 신호가 음성인지 여부를 판단하고, 등록된 사용자가 명령을 내리고 있는지 여부를 판단하고, 인가된 사용자가 발화하고 있는지 여부를 판단하며, 입력된 음성 신호가 미리 설정된 명령어를 포함하고 있는지 여부를 판단함으로써 버튼에 대한 수동 조작 없이도 사용자의 명령에 항시 응답이 가능한 음성 처리 방법을 제공할 수 있다. 각각의 과정들에 대해서는 이후에 도 5 내지 도 8을 통해 구체적으로 설명하도록 한다.

[0025] 도 3은 본 발명의 일 실시예에 따른 도 2의 음성 처리 방법에서 대상 신호를 강화하는 과정(220 단계)을 보다 구체적으로 설명하기 위한 흐름도로서, 다음과 같은 단계들을 포함한다.

[0026] 221 단계에서 음성 처리 장치로부터 발생되어 마이크로폰을 통해 입력되는 에코(echo) 신호를 제거한다. 일반적으로 마이크로폰이 스피커와 근접하게 배치되는 경우 스피커를 통해 출력된 소리가 마이크로폰에 입력되는 문제가 발생한다. 즉, 양방향 음성 입력과 출력이 동시에 이루어지는 상황에서, 스피커의 출력이 기기에 입력되어 재차 기기의 스피커의 출력으로 들리는 음향 반향(acoustic echo)이 발생하게 된다. 이러한 에코 신호는 정확한 음성 인식을 방해할 뿐만 아니라 사용자에게 큰 불편을 주기 때문에 제거되어야만 하는데, 이를 음향 반향 제거(AEC; acoustic echo cancellation)라고 한다. AEC가 이루어지는 과정을 간단하게 설명하면 다음과 같다.

[0027] 우선, 마이크로폰을 통해 사용자의 음성 및 간섭 잡음 이외에 스피커로부터 방사되는 출력 사운드가 포함된 혼합 사운드가 입력되는 것으로 가정한다. 221 단계에서는 스피커에 인가되는 출력 신호를 인자로 입력받아 특정 필터를 사용하여 마이크로폰을 통해 입력된 음원 신호로부터 스피커의 출력 신호를 제거한다. 이러한 필터는 시간의 흐름에 따라 지속적으로 스피커에 인가되는 출력 신호를 피드백(feed-back)받아 음원 신호에 포함된 음향 방향(acoustic echo)을 제거할 수 있는 적응 필터(adaptive filter)로 구성될 수 있다.

[0028] 이러한 AEC 방법에는 LMS(least mean square), NLMS(normalised least mean square), RLS(recursive least square) 등과 같은 다양한 알고리즘이 소개되어 있으며, 이상과 같은 방법들을 이용하여 AEC를 구현하는 방법은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 용이하게 파악이 가능한 것으로 여기에서는 자세한 설명을 생략한다.

[0029] 이제, 혼합 신호들로부터 에코 신호가 제거되었다. 이하에서는 소리 신호가 발생하는 위치를 찾아서 특정한 위치의 소리를 강화하기 위한 방법을 통하여 다양한 방향에서 발생하는 신호 중 특정 사용자의 음성에 대한 인식률을 높이는 단계들이 제시된다.

[0030] 222 단계에서는 수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색한다. 음향 위치를 추적하는 방법은 SRP(steered response power)와 같은 조향된 빔 형성기(steered beamformer)를 이용하거나, MUSIC(multiple signal classification)과 같은 고해상도 스펙트럼 추정(high resolution spectral estimation)을 이용하거나, GCC(generalized cross correlation)과 같은 도착 지연 시간(time delay of arrival)을 이용하는 방법 등이 활용될 수 있다. 각각의 방법들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 용이하게 파악이 가능한 것으로 여기에서는 자세한 설명을 생략한다.

[0031] 223 단계에서는 222 단계를 통해 탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거한다. 이러한 과정을 통해 222 단계를 통해 탐색된 신호 중 특정 방향(사용자 위치를 의미한다.)을 제외한 다른 위치에서 발생된 소리를 제거함으로써 연산의 대상을 감소시킬 수 있고, 결과적으로 빠른 음성 인식이 가능하다. 구체적인 처리 방법은 이하에서 도 4를 참고하여 설명한다.

[0032] 도 4는 본 발명의 일 실시예에 따른 도 3의 음성 처리 방법에서 잡음 신호를 제거하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

[0033] 223a 단계에서 신호들에 대한 빔 패턴(beam pattern) 중에서 미리 설정된 음원 방향에 해당하는 메인 로브(main lobe)를 제외한 사이드 로브(side lobe)에 대한 제거 신호를 생성한다. 빔 패턴이라 함은 스피커 및 마이크로폰 등의 신호 입출력 장치에서 입사 또는 방사되는 전자기파의 전계강도(electric field strength)를 측정하여 그래프로 표시한 것을 말한다. 따라서, 그래프의 기준점으로부터 멀리 떨어져 있을수록 전계강도가 크다는 것을 의미하고, 이는 해당 방향으로 지향성을 갖는다는 것을 의미한다. 빔 패턴에서는 중앙의 메인 로브(main lobe)를 중심으로 좌우로 작고 가는 빔 패턴들이 나타나는 것을 볼 수 있다. 지향성이 강하게 나타나는 메인 로브를 제외한 작고 가는 빔 패턴들을 사이드 로브(sibe lobe)라고 하며, 이러한 사이드 로브는 방사 패턴에서의 불균일 방사 패턴으로서 나타나게 된다. 그 결과, 사이드 로브 또는 불균일 방사 패턴은 음향 기기에서의 지향성과 같은 음장 특성이 수렴하는 것을 방해하는 요인이 된다.

[0034] 즉, 223a에서는 미리 설정된 음원 방향에 대한 강화된 신호는 남기고, 그 외의 신호들을 잡음으로 간주하여 제거하는 것을 목표로 한다. 구체적으로, 신호에 독립적인 빔 형성 기법의 출력 신호에 존재하는 사이드 로브 신호를 적응적으로 제거함으로서 대상 신호에 대한 지향성을 향상시킬 수 있다. 음성 신호와 잡음이 포함된 혼합 신호가 입력되면 우선, 고정 필터(fixed filter)를 이용하여 고정 빔 형성(fixed beamforming)함으로써 신호를 증폭시킨다. 한편, 블록킹 매트릭스(blocking matrix)를 통해 혼합 신호로부터 잡음 신호만을 추출하여 이를 제거 신호로 한다.

[0035] 223b 단계에서는 혼합 신호와 223a 단계를 통해 생성된 제거 신호를 합성함으로써 잡음 신호를 제거한다. 즉, 223a 단계를 통해 생성된 제거 신호는 고정 빔 형성된 혼합 신호 내의 잡음 신호를 제거하기 위해 적응 필터(adaptive filter)를 이용하는데, 혼합 신호로부터 잔여 잡음을 완전히 제거하기 위해 목표 신호(잡음이 제거된 신호를 의미한다.)인 출력 신호와 입력 신호와의 관계를 전달 함수로 표현하여 적응 필터에 대한 필터 계수를 변경시킴으로써 적절한 제거 신호를 생성할 수 있다.

[0036] 이러한 적응 필터를 활용하기 위해 개방형 피드-포워드(open-loop feed-forward) 방식이 활용될 수 있을 것이다. 개방형 피드-포워드 방식이란, 신호 처리 분야에서 임의의 변위값을 가질 수 있는 구성 요소를 시스템을 구성하는 다수의 구성 요소들과의 관계를 통해 제어하는 방법 중 하나로서 알려져 있다. 여기서, 제어란 어떤 물리량의 상태를 제어 주체가 원하는 목적에 맞는 상태가 되도록 하는 것을 의미한다. 즉, 본 발명의 실시예에서의 제어란 신호 처리 장치가 사용자가 원하는 결과를 출력할 수 있도록 신호 처리 장치의 특정 구성(제거 신호 생성 장치가 될 것이다.)이 가질 수 있는 변수를 조절하는 것을 말한다. 이를 통해 시스템을 구성하는 일련의 과정이 측정과 수행이 반복되는 폐-루프(close-loop)를 갖지 않고, 제어값을 검출하여 결과를 변화시키는 예측 제어 방식을 활용할 수 있다.

[0037] 이상과 같은 일련의 과정을 통해 본 발명의 실시예는 음성 신호와 잡음이 포함된 혼합 신호로부터 대상 음성 신호를 추출하여 명령어를 인식함으로써 음성 인식의 성능을 향상시키고, 수동 조작 없이 대상 기기를 용이하게 조작, 제어할 수 있다.

[0038] 도 5는 본 발명의 일 실시예에 따른 도 2의 음성 처리 방법에서 명령어를 인식하는 과정을 보다 구체적으로 설명하기 위한 흐름도로서, 다음과 같은 단계들을 포함한다.

[0039] 231 단계에서는 이상의 과정들을 통해 강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출한다. 일반적으로 소리 신호 내에 포함된 음성 구간을 판단함에 있어서, 음성 전체에 관한 특징 추출이 곤란하다는 문제와 신호 손상(signal corruption) 및 오픈-셋(open-set) 문제가 존재하는데, 본 실시예에서는 이러한 문제점을 해결하기 위해, 음성 인식 과정에 있어서 잡음 왜곡이 발생한 입력 음향 신호로부터 특징을 추출하는 과정 없이, 학습 단계를 통해 사전 학습된 잡음없는 모음 신호로부터 스펙트럼 특징을 추출 및 저장하고, 저장된 스펙트럼 특징을 이용하여 모음 신호 및 잡음 왜곡이 발생한 입력 음향 신호 간의 매칭을 수행하도록 한다. 그 이유는, 기존 기술의 신호 손상 및 오픈-셋 문제가 바로 잡음에 의해 왜곡된 입력 음향 신호로부터 추출된 특징을 이용하여 매칭을 수행함으로써 초래되었기 때문이다. 따라서, 본 발명의 실시예는 모음의 스펙트럼 특징만을 추

출하여 모음 신호 및 입력 음향 신호 간 매칭을 수행함으로써 유사도를 측정할 수 있다.

[0040] 이를 위해 미리 학습 단계를 통해 모음의 주파수 스펙트럼에서 특징 피크(characteristic peak)가 위치한 피크 대역(peak band)을 나타내는 모음 특징 정보를 저장한다. 모음의 스펙트럼에서 특징 피크를 추출함에 있어서, 다양한 형태의 추출 방식을 적용할 수 있는데, 본 실시예에서는 연산의 단순화를 위해 모음의 스펙트럼 피크들 중 미리 결정된 문턱값(threshold)보다 큰 에너지를 지니는 피크를 특징 피크로 추출할 수 있다. 인간의 음성(voice 또는 speech)은 크게 자음(consonant)과 모음(vowel)으로 구성된다. 모음 스펙트럼 상에 특징 피크들이 존재하는 피크 대역은 주변 대역에 비해 현저히 높은 에너지를 지니기 때문에 잡음에 의한 왜곡에도 쉽게 사라지지 않는다. 따라서, 인간의 음성 중 모음 부분은 다양한 잡음이 발생하는 실제 생활 환경에서도 검출이 용이하며, 이러한 모음 부분만 검출할 수 있으면 모음 앞의 수십 내지 수백 ms를 함께 음성으로 간주하는 방식을 이용하여 자음 및 모음 전체를 검출할 수 있게 된다.

[0041] 231 단계를 도 6을 참조하여 보다 구체적으로 설명한다. 도 6은 본 발명의 일 실시예에 따른 도 5의 음성 처리 방법에서 음성 구간을 검출하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다.

[0042] 우선, 231a 단계에서는 모음의 스펙트럼에서 특징 피크가 위치한 피크 대역을 나타내는 모음 특징 정보를 미리 저장한다.

[0043] 본 실시예에서는 모음의 전체 스펙트럼 대역을 일정 개수의 단위 대역으로 구별하여, 모음의 스펙트럼에서 피크 대역에 해당하는 단위 대역을 1로 나타내고, 피크 대역 이외의 대역인 밸리 대역(valley band)에 해당하는 단위 대역을 0으로 나타냄으로써 모음 특징 정보를 생성한다. 예컨대, 모음 특징 정보는 피크 특징 벡터(peak signature vector) 형태로 생성될 수 있다. 즉, 모음 신호에 관한 학습 데이터에 DFT(Discrete Fourier Transform)를 적용하여 얻어진 N 차원(N dimension)의 평균 스펙트럼으로부터 문턱값 이상의 특징 피크를 추출하면, 특징 피크가 위치한 피크 대역(peak band)에 해당하는 각각의 차원에 1을 할당하고, 피크 대역 이외의 대역인 밸리 대역(valley band)에 해당하는 각각의 차원에 0을 할당하여 N 차원의 피크 특징 벡터(peak signature vector)를 생성할 수 있다.

[0044] 다음으로, 231b 단계에서는 앞서 강화된 대상 신호들의 스펙트럼에서 231a 단계를 통해 저장된 모음 특징 정보가 나타내는 피크 대역에 대응하는 대응 대역 및 대응 대역을 제외한 비대응 대역의 평균 에너지를 이용하여 강화된 대상 신호들 중 음성에 해당하는 구간을 검출한다. 즉, 강화된 대상 신호들의 스펙트럼 상에서 이상에서 저장된 모음 특징 정보가 나타내는 피크 대역에 대응하는 대응 대역(relevant band) 및 대응 대역을 제외한 비대응 대역(irrelevant band)의 평균 에너지(average energy)를 이용하여 대상 신호들이 음성에 해당하는지 여부를 판단함으로써 음성 구간을 검출한다. 구체적으로, 대응 대역의 평균 에너지와 비대응 대역의 평균 에너지의 차이를 산출함으로써 이를 미리 결정된 임계값과 비교하고, 산출된 값이 임계값보다 큰 경우, 입력 음향을 모음을 포함하는 음성으로 판단하여 음성 구간으로 간주하고, 산출된 값이 임계값 이하인 경우 입력 음향을 잡음 또는 비음성 음향을 판단할 수 있다.

[0045] 이상의 과정을 통해 이제 음성 구간이 검출되었다. 다시 도 5로 돌아와서, 232 단계로 진행한다.

[0046] 232 단계에서는 231 단계를 통해 검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정한다. 이 과정은 음성이 존재할 때, 사용자가 제품에 명령을 내리는 것을 판단하고 그 대상을 한정하기 위함이다. 이러한 사용자를 한정하기 위해서는 다양한 방법들이 활용될 수 있으나, 이하에서는 도 7a 및 도 7b를 참조하여 2가지 실시예를 제시하도록 하겠다.

[0047] 첫째, 사용자를 음성 인식 대상으로 선택적으로 결정함에 있어서, 안면 인식을 이용하여 해당 위치의 사용자가 명령을 내리는 것인지 여부를 판단할 수 있다.

[0048] 구체적으로 도 7a에서, 232a 단계에서는 안면 인식을 이용하여 사용자가 상기 음성 처리 장치를 지향하고 있는지를 판단하고, 232b 단계에서는 안면 인식 결과에 기초하여 사용자가 미리 등록된 사용자인지 여부를 검사하며, 마지막으로 232c 단계에서는 232a 단계의 판단 결과 및 232b 단계의 검사 결과를 모두 만족하는 사용자를 상기 음성 인식 대상으로 결정한다. 이상과 같은 일련의 동작을 통해 해당 사용자가 이전에 인가된 사용자인지를 판단하여 권한 및 제품 동작의 정확성을 높일 수 있다. 안면 인식에 대한 실천적인 구현예와 설명은 본 발명의 본질을 벗어나는 것으로 여기서는 자세한 설명을 생략한다.

[0049] 둘째, 사용자를 음성 인식 대상으로 선택적으로 결정함에 있어서, 은닉 마코브 모델(hidden Markov models, HMM)을 이용하여 유효한 사용자인지 여부를 판단할 수 있다.

[0050] 구체적으로 도 7b에서 232d 단계에서는 사용자의 음성 신호 및 은닉 마코브 모델을 이용한 화자 인식 결과에 기초하여 사용자가 미리 등록된 사용자인지 여부를 검사하고, 232e 단계에서는 232d 단계의 검사 결과를 만족하는 사용자를 음성 인식 대상으로 결정한다.

[0051] 특히, 본 발명의 실시예에서는 부분 공간 분배 클러스터링 은닉 마코브 모델(Subspace Distribution Clustering Hidden Markov Model, SDCHMM)을 사용할 수 있다. SDCHMM은 전체 다변수 가우시안(multivariate Gaussian)들을 부분 공간(subspace)이라는 특징 공간(feature space)들로 나눈 뒤, 각 부분 공간에 속한 분포들을 양자화하여 양자화된 부분 공간 분포 프로토타입 즉, 코드워드(codeword)들을 조합하여 원래의 전체 공간(full space) 분포를 표현한다. 즉, 원래의 전체 공간 분포와 가장 유사한 부분 공간 프로토타입들의 조합을 찾아서 그것들을 링크한다. 이와 같이 SDCHMM은 적은 수의 부분공간 코드워드 (subspace codeword)를 이용하여 원래의 전체 공간 분포를 표현할 수 있다.

[0052] 구체적으로 부분 공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 전체 공간에 대한 은닉 마코브 모델 (HMM)로 변환하고, 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환한 다음, 최대 우도 선형 스펙트럼 변환을 이용하여 이상의 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응시킴으로써 고속으로 사용자를 인식할 수 있다.

[0053] 이상의 과정을 통해 이제 사용자가 결정되었다. 다시 도 5로 돌아와서, 233 단계로 진행한다.

[0054] 233 단계에서는 232 단계를 통해 결정된 음성 인식 대상의 음성 구간으로부터 미리 설정된 명령어를 탐지한다. 키워드 인식 과정을 통해 사람의 다양한 언어 중 대상 기기의 동작 과정에 필요한 언어, 즉 명령어인지 여부를 판별한다. 그로 인해 음성 처리 장치가 항시 소리 신호 중 명령어를 입력받을 수 있는 대기 상태를 유지하면서, 오작동하는 확률을 줄일 수 있으며, 사용자는 특정한 동작(버튼의 수동 입력 등이 될 수 있다.) 없이도 해당 기기에 적절한 명령을 내릴 수 있다.

[0055] 도 8은 본 발명의 일 실시예에 따른 도 5의 음성 처리 방법에서 미리 설정된 명령어를 탐지하는 과정을 보다 구체적으로 설명하기 위한 흐름도이다. 이러한 과정은 인식된 음성 명령어를 이용하여 문장과 제품의 명령어와의 대조(matching) 작업을 통해 일치하는 명령군이 존재하는지 여부를 판단함으로써 이루어진다.

[0056] 우선, 233a 단계에서 사용자의 의해 발음될 수 있는 음절의 대표 발음들로 구성되는 음절 모델과 명령어 세트를 미리 저장한다. 여기서, 음절 모델은 사용자에 의해 발음될 수 있는 음절들의 대표 발음들로 구성된다. 음절 모델은 유사한 발음의 이중 모음들을 갖는 음절들에 대해 한 가지의 음절 발음만을 저장하고, 유사한 발음의 받침을 갖는 음절들에 대해 하나의 공통된 음절 발음만을 저장하도록 구성될 수 있다. 한글처럼 표음문자인 언어는 한 음절이 한 글자로 표기가 가능하다. 따라서, 음절 단위로 발화된 음성은 연속 발음에서 발생하는 조음현상의 발생을 억제하여 비교적 정확한 발음 입력을 가능하게 한다. 한글 완성형 코드는 2350자이고 구별할 수 있는 대표 발음들로 묶으면 1000개 내외로 적은 양의 음절 모델을 생성할 수 있다. 또한, 연속 발음된 음성도 가장 유사한 음절 후보 리스트를 탐색하고 언어 모델을 통해 생성 가능한 어절 및 문장을 구성하면 비교적 적은 음절 모델을 통해서도 정확한 음성 인식이 가능하다.

[0057] 233b 단계에서는 앞서 결정된 음성 인식 대상의 음성 구간으로부터 각 음절의 특징을 추출한다. 다음으로, 233c 단계에서는 233b 단계를 통해 추출된 특징을 233a 단계를 통해 저장된 음절 모델과 비교하여 적어도 하나 이상의 음절 후보를 생성한다. 즉, 233a 단계를 통해 생성된 음절 모델을 이용하여 복수의 음절 후보로 조합 가능한 어절들의 확률을 연산하고, 어절들의 확률에 따라 어절들이 연결된 복수의 문장 후보를 생성한다. 그런 다음, 233d 단계에서는 233c 단계를 통해 생성된 음절 후보로부터 조합 가능한 어절들 중 저장된 명령어 세트에 포함되는 명령어를 탐지한다.

[0058] 이상에서 설명한 본 발명의 실시예들에 따르면 음성 인식을 위한 음성 입력 과정에서 모음 특징을 이용하여 음성 구간을 검출함으로써 별도의 번거로운 수동 조작(음성 입력의 시작과 끝(Push to Talk)을 지정하는 조작을 의미한다.) 없이도 연속적인 음성 신호에서 명령어를 추출, 인식할 수 있다. 특히, 이상의 실시예들은 제한적 환경에서 사용되었던 음성 인터페이스를 다양한 환경 및 제품에 적용이 가능하고, 높은 인식 성능을 나타낸다.

[0059] 도 9는 본 발명의 일 실시예에 따른 잡음 환경에서의 음성 처리 장치(900)를 도시한 블록도로서, 수신부(10), 신호 강화부(20) 및 음성 명령어 인식부(30)를 포함한다. 나아가, 본 발명의 실시예는 인식 향상부(40)를 선택적으로 포함할 수 있다. 이하에서 각각의 구성 요소를 차례로 설명하되, 이미 이상의 실시예들을 통해 설명된 내용과 중복되는 내용에 대해서는 약술하도록 한다.

[0060] 수신부(10)는 마이크로폰을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신한다. 이러한 수신부(10)는 도 2의 210 단계에 대응한다.

[0061] 신호 강화부(20)는 수신부(10)를 통해 수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화한다. 이러한 신호 강화부(20)는 도 2의 220 단계에 대응하는 것으로, 도 10을 통해 세부 구성 요소를 설명한다.

[0062] 신호 강화부(20)는 음성 처리 장치로부터 발생되어 마이크로폰을 통해 입력되는 에코 신호를 제거하는 에코 제거부(21), 수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색하는 방향 탐색부(22) 및 탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거하는 잡음 제거부(23)를 포함한다. 이상의 세부 구성들(21, 22, 23)은 각각 도 3의 221 단계 내지 223 단계에 대응하는 것으로 여기서는 구체적인 설명을 생략한다.

[0063] 음성 명령어 인식부(30)는 음성 처리 장치를 지향하는 사용자만을 대상으로 앞서 신호 강화부(20)를 통해 강화된 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식한다. 이러한 음성 명령어 인식부(30)는 도 2의 230 단계에 대응하는 것으로, 도 11을 통해 세부 구성 요소를 설명한다.

[0064] 음성 명령어 인식부(30)는 강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출하는 음성 검출부(31), 검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정하는 대상 결정부(32) 및 결정된 음성 인식 대상의 음성 구간으로부터 미리 설정된 명령어를 탐지하는 명령어 탐지부(33)를 포함한다. 이상의 세부 구성들(31, 32, 33)은 각각 도 5의 231 단계 내지 233 단계에 대응하는 것으로, 여기서는 구체적인 설명을 생략한다.

[0065] 한편, 인식 향상부(40)는 이상과 같은 과정을 통해 얻어진 정보의 음성 인식을 향상시키기 위한 구성에 해당한다. 이러한 인식 향상부(40)는 다양한 음성 인식 모델들이 활용될 수 있으나, 본 발명의 실시예에서는 진화 학습을 이용한 화자 적응 방법이 활용 가능하다.

[0066] 음성 인식 시스템이 음성 인식을 수행하고 있는 인식 모드에서, 신속한 환경 적응이 가능한 특징 변환을 수행하도록 하고, 음성 인식 시스템이 음성 인식을 수행하지 않고 있는 대기 모드에서, 미리 저장해 둔 충분한 양의 음성 데이터를 적응 데이터로 이용하여 모델 변환을 수행하도록 한다. 예를 들어, 이러한 특징을 변환하는 과정은 최대 우도 기법(maximum likelihood method)을 이용하여 환경 파라미터를 미리 결정함으로써 수행될 수 있을 것이다. 즉, 본 실시예는 화자 적응을 위한 진화 학습(evolutional learning) 과정으로서 특징 변환 및 모델 변환을 반복적으로 수행하도록 하여 축적된 적응 데이터의 양과 무관하게 안정적이고 높은 인식성능을 보장한다.

[0067] 화자 적응 기법은 화자 독립 음향모델을 추정하는 과정과 이를 통해 추정된 화자 독립 음향모델을 화자 종속 음향모델로 재추정하는 과정으로 이루어진다. 즉, 불특정 다수의 화자로부터 얻어진 많은 양의 음성데이터에서 MFCC(Mel-Frequency Cepstral Coefficient)와 같은 다차원의 특징 벡터를 추출하고, 추출된 특징 벡터들에 대해 EM(Expectaion-Maximization) 기법 등을 이용하여 HMM(Hidden Markov Model)을 추정한다. 이렇게 추정된 HMM은 불특정 다수의 화자에 대한 음향모델로서 화자 독립적인 특성을 지닌다. 또한, 본 실시예는 이미 추정된 화자 및 환경 독립 음향모델과 특정 화자(사용자)로부터 획득되는 상대적으로 적은 양의 음성 데이터를 이용하여 화자 및 환경 종속 음향모델을 추정할 수 있다.

[0068] 한편, 본 발명은 컴퓨터로 읽을 수 있는 기록 매체에 컴퓨터가 읽을 수 있는 코드로 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록 매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다.

[0069] 컴퓨터가 읽을 수 있는 기록 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현하는 것을 포함한다. 또한, 컴퓨터가 읽을 수 있는 기록 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다. 그리고 본 발명을 구현하기 위한 기능적인(functional) 프로그램, 코드 및 코드 세그먼트들은 본 발명이 속하는 기술 분야의 프로그래머들에 의하여 용이하게 추론될 수 있다.

[0070] 이상에서 본 발명에 대하여 그 다양한 실시예들을 중심으로 살펴보았다. 본 발명에 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있

는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0071] 110, 120, 130, 140 : 잡음 환경에서의 다수의 사용자들

900 : 음성 처리 장치

10 : 수신부

20 : 신호 강화부 21 : 에코 제거부

22 : 방향 탐색부 23 : 잡음 제거부

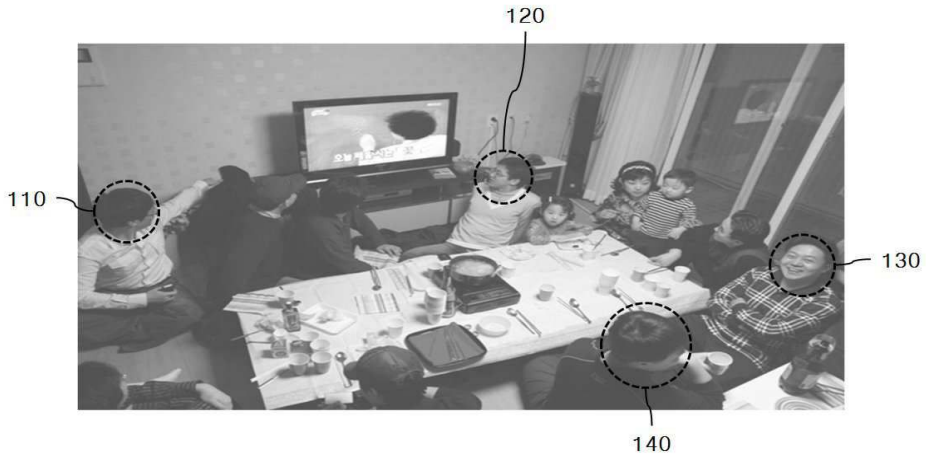
30 : 음성 명령어 인식부 31 : 음성 검출부

32 : 대상 결정부 33 : 명령어 탐지부

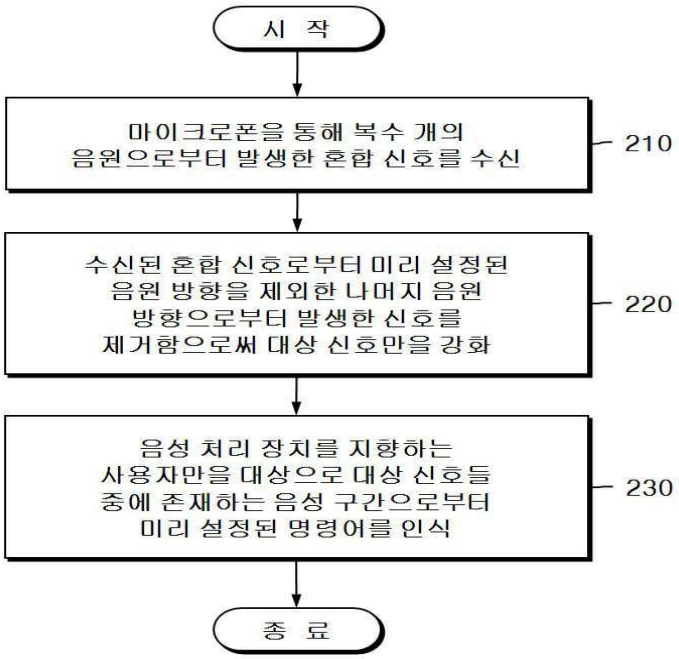
40 : 인식 향상부

도면

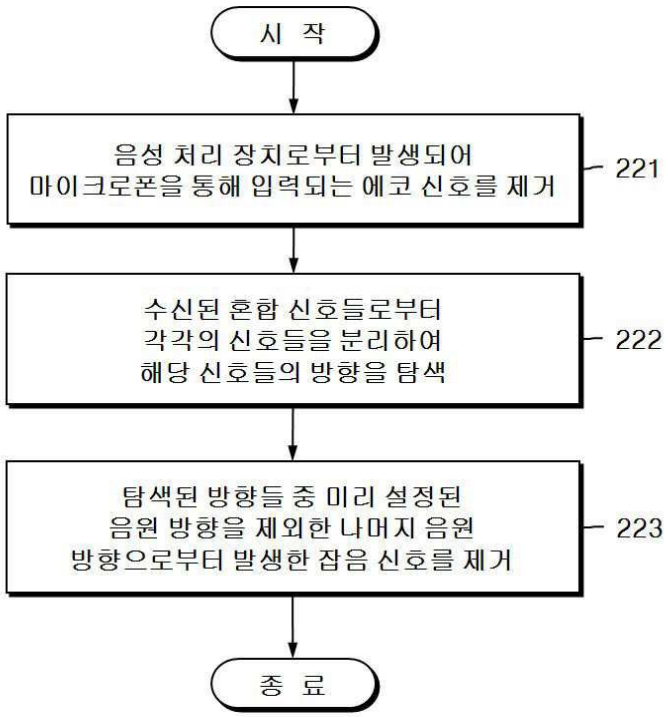
도면1



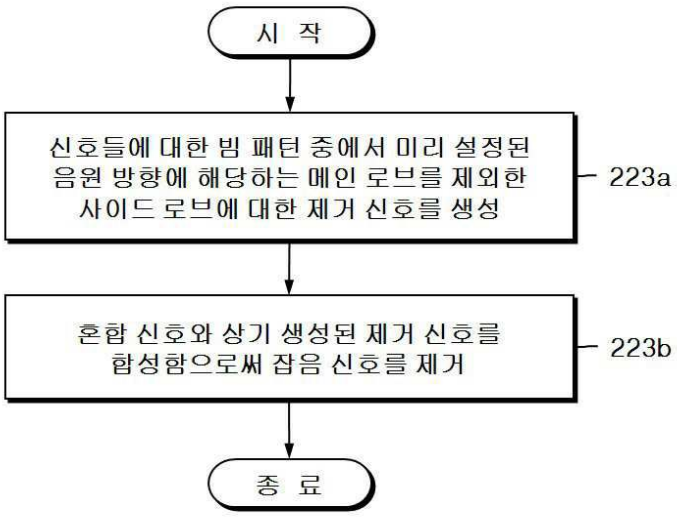
도면2


시 작
마이크로폰을 통해 복수 개의 음원으로부터 발생한 혼합 신호를 수신
210
수신된 혼합 신호로부터 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 신호를 제거함으로써 대상 신호만을 강화
220
음성 처리 장치를 지향하는 사용자만을 대상으로 대상 신호들 중에 존재하는 음성 구간으로부터 미리 설정된 명령어를 인식
230
종 료

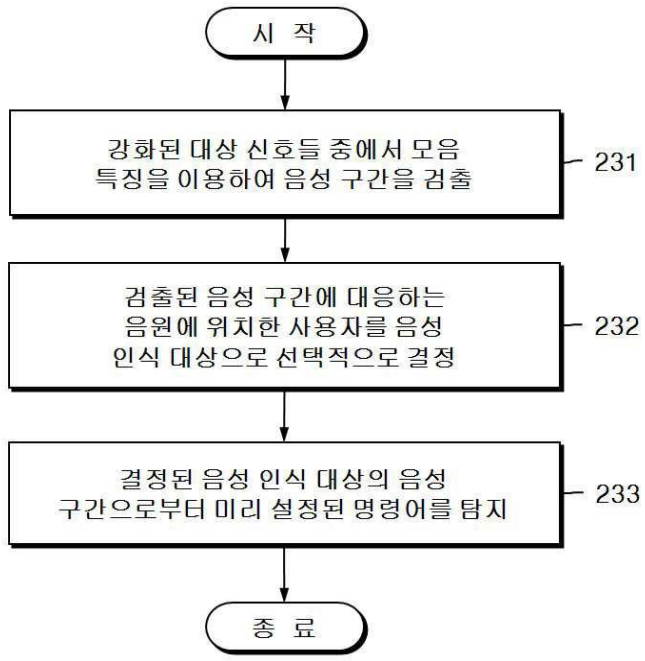
도면3


시 작
음성 처리 장치로부터 발생되어 마이크로폰을 통해 입력되는 에코 신호를 제거
221
수신된 혼합 신호들로부터 각각의 신호들을 분리하여 해당 신호들의 방향을 탐색
222
탐색된 방향들 중 미리 설정된 음원 방향을 제외한 나머지 음원 방향으로부터 발생한 잡음 신호를 제거
223
종 료

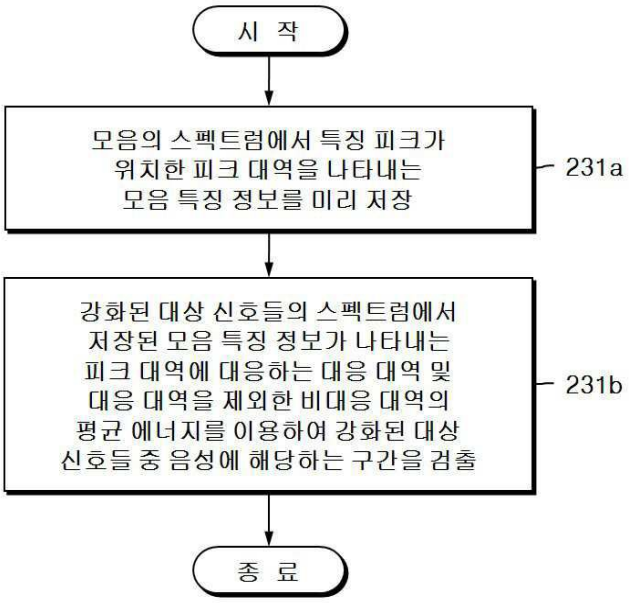
도면4


시 작
신호들에 대한 빔 패턴 중에서 미리 설정된 음원 방향에 해당하는 메인 로브를 제외한 사이드 로브에 대한 제거 신호를 생성
223a
혼합 신호와 상기 생성된 제거 신호를 합성함으로써 잡음 신호를 제거
223b
종 료

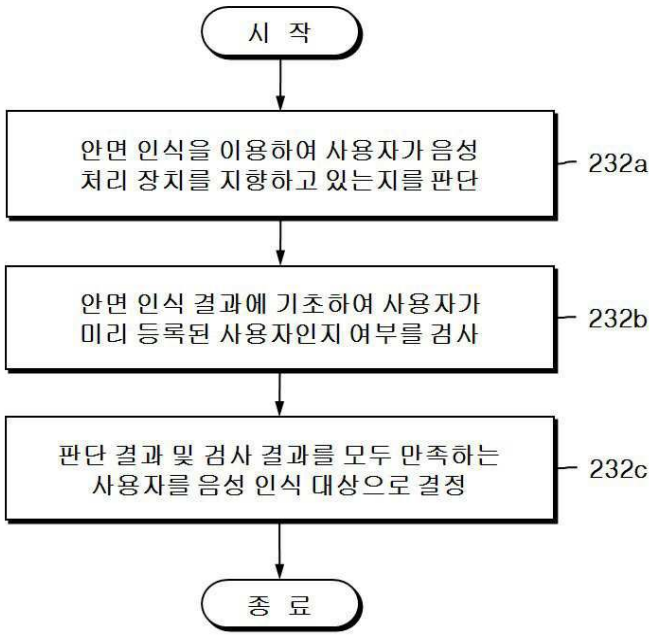
도면5


시 작
강화된 대상 신호들 중에서 모음 특징을 이용하여 음성 구간을 검출
231
검출된 음성 구간에 대응하는 음원에 위치한 사용자를 음성 인식 대상으로 선택적으로 결정
232
결정된 음성 인식 대상의 음성 구간으로부터 미리 설정된 명령어를 탐지
233
종 료

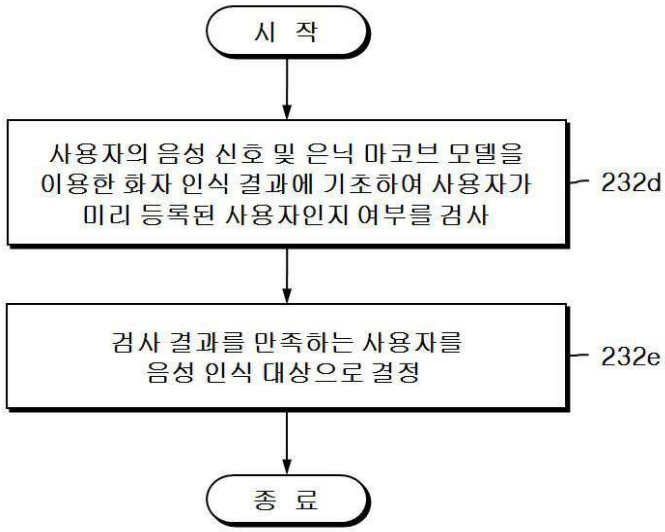
도면6


시 작
모음의 스펙트럼에서 특징 피크가 위치한 피크 대역을 나타내는 모음 특징 정보를 미리 저장
231a
강화된 대상 신호들의 스펙트럼에서 저장된 모음 특징 정보가 나타내는 피크 대역에 대응하는 대응 대역 및 대응 대역을 제외한 비대응 대역의 평균 에너지를 이용하여 강화된 대상 신호들 중 음성에 해당하는 구간을 검출
231b
종 료

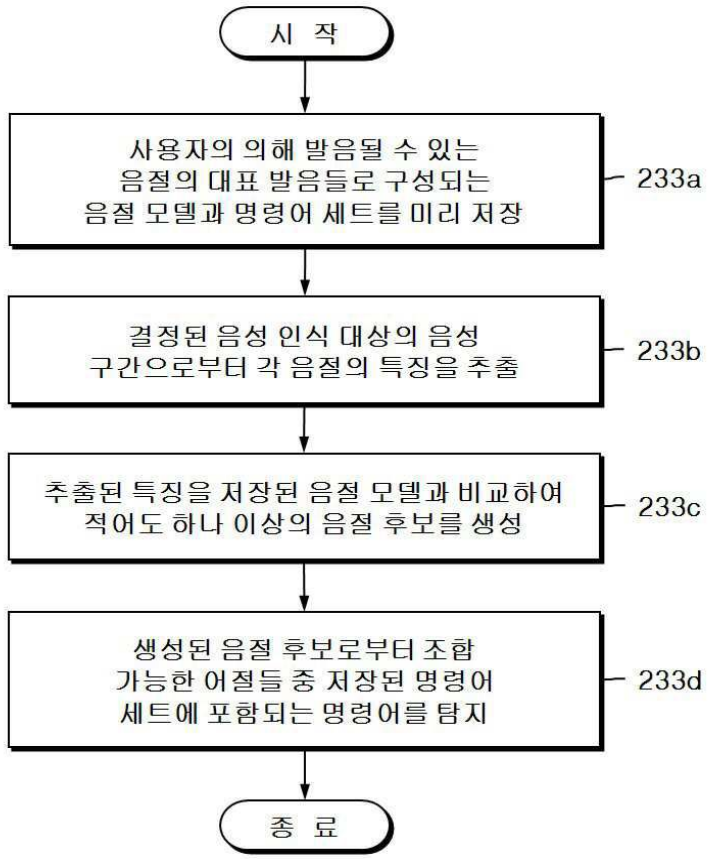
도면7a


시 작
안면 인식을 이용하여 사용자가 음성 처리 장치를 지향하고 있는지를 판단
232a
안면 인식 결과에 기초하여 사용자가 미리 등록된 사용자인지 여부를 검사
232b
판단 결과 및 검사 결과를 모두 만족하는 사용자를 음성 인식 대상으로 결정
232c
종 료

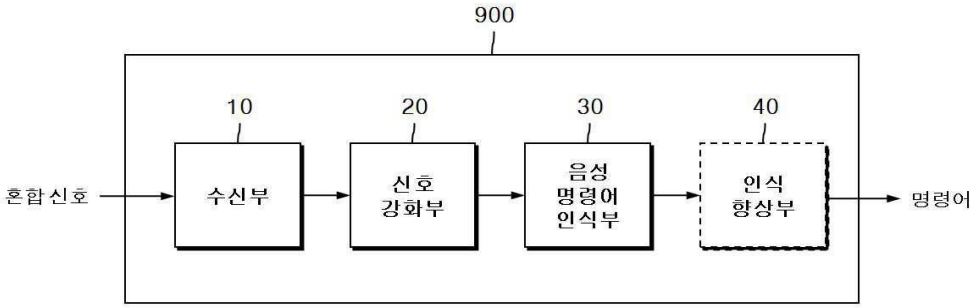
도면7b


시 작
사용자의 음성 신호 및 은닉 마코브 모델을 이용한 화자 인식 결과에 기초하여 사용자가 미리 등록된 사용자인지 여부를 검사
232d
검사 결과를 만족하는 사용자를 음성 인식 대상으로 결정
232e
종 료

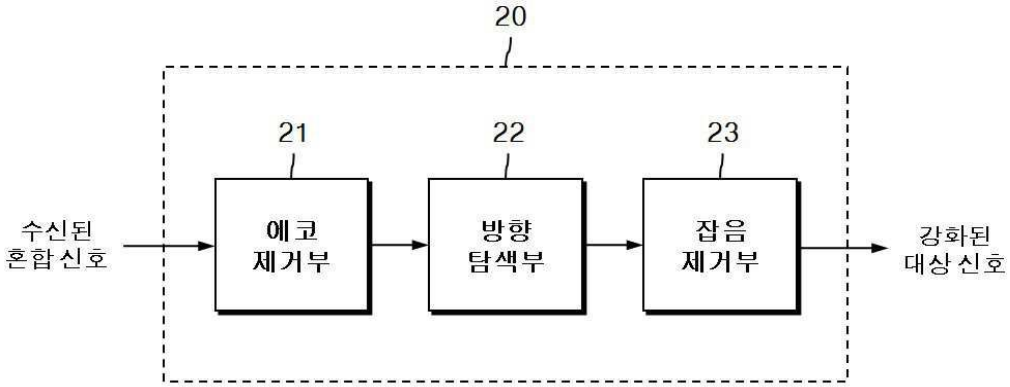
도면8


시 작
사용자의 의해 발음될 수 있는 음절의 대표 발음들로 구성되는 음절 모델과 명령어 세트를 미리 저장
233a
결정된 음성 인식 대상의 음성 구간으로부터 각 음절의 특징을 추출
233b
추출된 특징을 저장된 음절 모델과 비교하여 적어도 하나 이상의 음절 후보를 생성
233c
생성된 음절 후보로부터 조합 가능한 어절들 중 저장된 명령어 세트에 포함되는 명령어를 탐지
233d
종 료

도면9



도면10



도면11

