(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년04월21일
(11) 등록번호 10-2102387
(24) 등록일자 2020년04월13일

(51) 국제특허분류(Int. Cl.)
G10L 25/87 (2013.01) *G10L 17/02* (2013.01)
G10L 17/04 (2013.01) *G10L 17/06* (2013.01)
(52) CPC특허분류
G10L 25/87 (2013.01)
G10L 17/02 (2013.01)
(21) 출원번호 10-2018-0129863
(22) 출원일자 2018년10월29일
심사청구일자 2018년10월29일
(56) 선행기술조사문헌
KR1019990038741 A*
KR1020160088446 A*
KR1020170045123 A*
KR1020180023702 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 사운드잇
서울특별시 서초구 사임당로 180 , 407호(서초동, 보원빌딩)
(72) 발명자
김도훈
서울특별시 강동구 구천면로42길 73-6, 402호 (천호동, 해피리움)
최인정
경기도 화성시 영통로27번길 20, 406동 1804호 (반월동, 신영통 현대타운)
(74) 대리인
리앤목특허법인

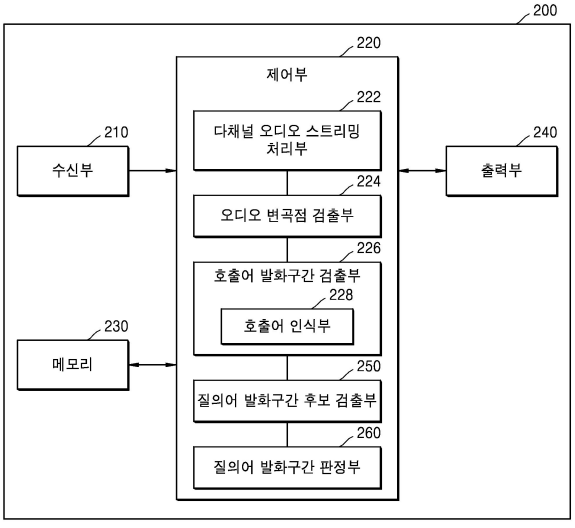
전체 청구항 수 : 총 8 항 심사관 : 이선희

(54) 발명의 명칭 **다채널오디오스트리밍에서 화자의 발화구간을 검출하는 방법 및 시스템**

(57) 요 약

본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법은 오디오 변곡점을 이용하여 다채널 오디오 환경하에서 호출어 인식에 필요한 계산량을 줄이며, 오디오 변곡점 지점에서만 호출어 발화구간과 질의어 발화구간을 검출하고, 호출어 발화구간과 질의어 발화구간 각각에서 검출한 화자의 유사도를 검출하여 화자 인식의 정확도를 높일 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G10L 17/04 (2013.01)
G10L 17/06 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 S2534002
부처명 중소벤처기업부
연구관리전문기관 중소기업기술정보진흥원
연구사업명 2017년 창업성장기술개발사업
연구과제명 다채널 음성처리 및 화자식별 기반 회의용 실시간 화자분할 기술 개발
기 여 율 1/1
주관기관 주식회사 사운드잇
연구기간 2017.10.30 ~ 2018.10.29

명 세 서

청구범위

청구항 1

오디오변곡점검출부에서 화자의 음성발화와 주변 소리가 포함된 다채널오디오스트리밍입력에서 음원 속성이 변경되거나 또는 화자가 변경되는 지점을 나타내는 적어도 하나의 오디오 변곡점을 검출하는 단계;

호출어발화구간검출부에서 상기 다채널오디오스트리밍입력 중 상기 검출된 적어도 하나의 오디오 변곡점에 대해서만 호출어 모델과의 유사도를 측정하여 화자의 호출어발화구간을 검출하는 단계;

호출어인식부에서 상기 호출어발화구간 내에서 기지정된 호출어를 인식하는 단계;

질의어발화구간후보검출부에서 상기 다채널오디오스트리밍입력에서 상기 검출된 호출어발화구간 이후에 뒤따라오는 구간에서 검출된 적어도 하나의 오디오 변곡점의 조합을 기초로 적어도 하나의 질의어발화구간후보를 모두 검출하는 단계;

질의어발화구간판정부에서 상기 검출된 호출어발화구간과 상기 검출된 적어도 하나의 질의어발화구간후보 각각을 비교하여 화자유사도를 측정하는 단계로, 상기 검출된 적어도 하나의 오디오 변곡점이 복수 개인 경우, 제 1 시간에 검출된 오디오 변곡점과 시간순으로 상기 제 1 시간 이후 최초로 검출된 오디오 변곡점 간에 생성된 질의어발화구간후보에서 검출된 상기 화자유사도가 기설정된 값을 초과하지 않는 경우, 상기 제 1 시간에 검출된 오디오 변곡점을 기초로 생성된 상기 질의어발화구간후보에 대해서는 상기 화자유사도를 측정하지 않고, 상기 제 1 시간 이후 최초로 검출된 오디오 변곡점을 기초로 생성된 질의어발화구간후보에서 상기 화자유사도를 측정하는, 화자유사도 측정단계;및

상기 질의어발화구간후보판정부는 상기 적어도 하나의 질의어발화구간후보 중 상기 화자유사도가 기설정된 값을 초과하는 질의어발화구간후보들을 선택하고, 선택된 질의어발화구간후보들 중 최장의 발화구간을 질의어발화구간으로 판정하는 단계;를 포함하고, 이 경우 상기 호출어발화구간은 상기 오디오 변곡점을 검출하는 단계에서 검출된 적어도 하나의 오디오 변곡점으로 판정된 지점들 중에서만 시작지점과 끝지점을 선정하여 검출되는 것을 특징으로 하는 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법.

청구항 2

제 1 항에 있어서,

프로세서로 구현되는 제어부에서 상기 다채널오디오스트리밍입력을 전처리(pre-processing)하여 특징벡터를 추출하고, 추출한 특징벡터를 이용하여 파악한 피치정보를 기초로 안정적인 피치 지속길이(stable pitch duration)를 결정하며, 그리고 피치변곡점을 검출하고, 상기 적어도 하나의 오디오 변곡점을 검출하는 단계는 상기 제어부에서 검출된 피치변곡점에 대해서만 상기 적어도 하나의 오디오 변곡점을 검출하는 것을 특징으로 하는 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법.

청구항 3

제 2 항에 있어서, 상기 적어도 하나의 오디오 변곡점을 검출하는 단계는

상기 다채널오디오스트리밍입력의 전처리시 획득되는 도착시간지연특징에 기초한 음원방향 정보를 더 이용하는 것을 특징으로 하는 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법.

청구항 4

제 1 항에 있어서,

상기 호출어발화구간에서 인식된 화자가 기등록된 화자인 경우, 상기 질의어발화구간후보들에서 인식된 화자에 대해서는 텍스트독립형 성문모델과의 식별점수를 기준으로 기등록된 화자의 발화인지 여부를 판단하는 것을 특징으로 하는 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법.

청구항 5

화자발화구간을 검출하는 시스템으로서,

상기 시스템은

스피커, 빔포머, 메모리, 프로세서;

주변환경에서 캡처한 사운드에서 적어도 하나 이상의 오디오스트리밍입력을 생성하는 복수의 마이크로폰;및

화자 모델(speaker model) 및 상기 프로세서에서 실행되는 컴퓨터로 실행가능한 명령을 저장하는 메모리;를 포함하고,

상기 프로세서는

다채널오디오스트리밍입력에서 추출한 특징벡터를 기초로 피치정보를 파악하여 피치변곡점을 검출하는 단계;

상기 피치변곡점에 대해서만 상기 다채널오디오스트리밍입력의 적어도 하나의 오디오 변곡점을 검출하는 단계;

상기 적어도 하나의 오디오 변곡점 지점을 기초로 호출어 발화구간과 질의어 발화구간을 검출하는 단계;

상기 호출어 발화구간에서 식별한 화자가 상기 메모리에 기등록된 화자인지를 판단하는 단계;및

상기 호출어 발화구간의 화자와 상기 질의어 발화구간의 화자가 일치하는지를 판단하여, 일치하는 경우 해당 화자의 질의어에 대한 응답을 제공하는 단계;를 실행하는 명령어를 저장하고,

상기 질의어 발화구간을 검출하는 단계는

검출된 호출어 발화구간과 적어도 하나의 질의어발화구간후보 각각을 비교하여 화자유사도를 측정하는 단계를 더 포함하고, 상기 화자유사도를 측정하는 단계는 상기 적어도 하나의 오디오 변곡점이 복수 개인 경우, 제 1 시간에 검출된 오디오 변곡점과 시간순으로 상기 제 1 시간 이후 최초로 검출된 오디오 변곡점 간에 생성된 질의어발화구간후보에서 검출된 화자유사도가 기설정된 값을 초과하지 않는 경우, 상기 제 1 시간에 검출된 오디오 변곡점을 기초로 생성된 상기 질의어발화구간후보에 대해서는 상기 화자유사도를 측정하지 않고, 상기 제 1 시간 이후 최초로 검출된 오디오 변곡점을 기초로 생성된 질의어발화구간후보에서 상기 화자유사도를 측정하여 상기 화자유사도가 기설정된 값을 초과하는 질의어발화구간후보들을 선택하고, 선택된 질의어발화구간후보들 중 최장의 발화구간을 질의어발화구간으로 검출하는 것을 특징으로 하는 시스템.

청구항 6

제 5 항에 있어서, 상기 프로세서는

상기 호출어 발화구간에서 식별한 화자가 상기 기등록된 화자인 경우, 상기 질의어 발화구간의 화자를 식별할 때 상기 기등록된 화자의 텍스트독립형 성문모델과의 식별점수를 기준으로 상기 호출어 발화구간의 화자와 일치하는지를 판단하는 단계;를 실행하는 명령어를 더 저장하는 것을 특징으로 하는 시스템.

청구항 7

제 5 항에 있어서, 상기 적어도 하나의 오디오 변곡점을 검출하는 단계는

상기 빔포머를 이용하여 파악한 음원방향 정보를 더 이용하여 상기 오디오 변곡점을 검출하는 것을 특징으로 하는 시스템.

청구항 8

제 5 항에 있어서,

상기 호출어 발화구간의 화자와 상기 질의어 발화구간의 화자가 일치하는지를 판단하기 위해 화자유사도를 측정하여 판단하는 것을 특징으로 하는 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 화자 인식에 관한 것이다. 보다 상세히, 다채널 오디오 생성 환경하에서 특정 화자의 발화를 인식하는 방법에 관한 것이다.

배 경 기 술

[0002] 최근에 인공지능 스피커의 보급이 활발해지고 있으며, 특히 인공지능 스피커에서 음성 입출력에 의한 질의응답 기능이 필수 기능으로 자리잡고 있다. 인공지능 스피커는 인공지능 스피커 자체에서 음악이 재생되는 환경 등에서는 반향 제거 기술 등을 적용하여 사용자 명령을 인식하고, 답변을 제공할 수 있다. 그러나, TV나 다른 오디오 기기에서 방송이나 음악이 재생되는 환경 하에서는 인공지능 스피커가 사용자의 음성을 인식하는데 있어 한계가 있다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) JP 2016-536626

발명의 내용

해결하려는 과제

[0004] 본 발명의 바람직한 일 실시예에서는 주위 다른 사람이 말하는 중이거나, TV 또는 다른 오디오 기기에서 방송이나 음악이 재생되는 환경하에서, 특정 화자의 음성발화 구간을 정확하게 검출하는 방법 및 장치를 제안하고자 한다.

[0005] 본 발명의 또 다른 바람직한 일 실시예에서는 일반적인 호출어 발화구간 검색 및 인식의 경우 음성구간이라고 판별되는 모든 순간마다 호출어의 시작과 끝이 가능하다고 가정하여 인식을 수행하게 되므로 특정화자의 호출어 발화가 없는 일반적인 유효 음향이 입력되는 순간에도 상당한 계산량이 소요되는 문제를 해결하고자 한다.

과제의 해결 수단

[0006] 본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법은 화자의 음성발화와 주변 소리가 포함된 다채널오디오스트리밍입력에서 음원 속성이 변경되거나 또는 화자가 변경되는 지점을 나타내는 적어도 하나의 오디오 변곡점을 검출하는 단계; 상기 다채널오디오스트리밍입력 중 상기 검출된 적어도 하나의 오디오 변곡점을 기준으로 호출어 모델과의 유사도를 측정하여 화자의 호출어발화구간을 검출하는 단계; 상기 다채널오디오스트리밍입력에서 상기 검출된 호출어 발화구간 이후에 뒤따라오는 구간에서 검출된 적어도 하나의 오디오 변곡점의 조합을 기초로 적어도 하나의 질의어발화구간후보를 모두 검출하는 단계; 상기 검출된 호출어발화구간과 상기 검출된 적어도 하나의 질의어발화구간후보 각각을 비교하여 화자유사도를 측정하는 단계;및 상기 적어도 하나의 질의어발화구간후보 중 상기 화자유사도가 기설정된 값을 초과하는 질의어발화구간후보들을 선택하고, 선택된 질의어발화구간후보들 중 최장의 발화구간을 질의어발화구간으로 판정하는 단계;를 포함하는 것을 특징으로 한다.

[0007] 본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법은 프로세서로 구현되는 제어부에서 상기 다채널오디오스트리밍입력을 전처리(pre-processing)하여 특징벡터를 추출하고, 추출한 특징벡터를 이용하여 파악한 피치정보를 기초로 안정적인 피치 지속길이(stable pitch duration)를 결정하며, 그리고 피치변곡점을 검출하고, 상기 적어도 하나의 오디오 변곡점을 검출하는 단계는 상기 제어부에서 검출된 피치변곡점에 대해서만 상기 적어도 하나의 오디오 변곡점을 검출하는 것을 특징으로 하는 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 것을 특징으로 한다.

[0008] 본 발명의 또 다른 바람직한 일 실시예로서, 화자발화구간을 검출하는 시스템에서 화자의 발화구간을 검출하는 방법으로서, 상기 시스템은 스피커, 빔포머, 메모리, 프로세서; 주변환경에서 캡처한 사운드에서 적어도 하나 이상의 오디오스트리밍입력을 생성하는 복수의 마이크로폰; 및 화자 모델(speaker model) 및 상기 프로세서에서 실행되는 컴퓨터로 실행가능한 명령을 저장하는 메모리;를 포함하고, 상기 프로세서는 다채널오디오스트리밍입력에서 추출한 특징벡터를 기초로 피치정보를 파악하여 피치변곡점을 검출하는 단계; 상기 피치변곡점에 대해서만 상기 다채널오디오스트리밍입력의 적어도 하나의 오디오 변곡점을 검출하는 단계;상기 적어도 하나의 오디오

변곡점 지점을 기초로 호출어 발화구간과 질의어 발화구간을 각각 검출하는 단계; 상기 호출어 발화구간에서 식별한 화자가 상기 메모리에 기등록된 화자인지를 판단하는 단계;및 상기 호출어 발화구간의 화자와 상기 질의어 발화구간의 화자가 일치하는지를 판단하여, 일치하는 경우 해당 화자의 질의어에 대한 응답을 제공하는 단계;를 실행하는 명령어를 저장하는 것을 특징으로 한다.

[0009] 본 발명의 또 다른 바람직한 일 실시예로서, 상기 프로세서는 상기 호출어 발화구간에서 식별한 화자가 상기 기등록된 화자인 경우, 상기 질의어 발화구간의 화자를 식별할 때 상기 기등록된 화자의 텍스트독립형 성문모델과의 식별점수를 기준으로 상기 호출어 발화구간의 화자와 일치하는지를 판단하는 단계;를 실행하는 명령어를 더 저장하는 것을 특징으로 한다.

[0010] 본 발명의 또 다른 바람직한 일 실시예로서, 상기 적어도 하나의 오디오 변곡점을 검출하는 단계는 상기 빔포머를 이용하여 파악한 음원방향 정보를 더 이용하여 상기 오디오 변곡점을 검출하는 것을 특징으로 한다.

[0011] 본 발명의 또 다른 바람직한 일 실시예로서, 상기 호출어 발화구간의 화자와 상기 질의어 발화구간의 화자가 일치하는지를 판단하기 위해 화자유사도를 측정하여 판단하는 것을 특징으로 한다.

발명의 효과

[0012] 본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 방법은 주위 다른 사람이 말하는 중이거나, TV 또는 다른 오디오 기기에서 방송이나 음악이 재생되는 환경 하에서서도 화자가 인공지능 스피커나 스마트 TV에 음성 명령을 내릴 때, 오디오 변곡점 검출에 의해 호출어와 질의어 발화구간을 정교하게 검출하는 효과가 있다.

[0013] 또한, 본 발명의 또 다른 바람직한 일 실시예에서는 다채널 음성처리에 의한 음원 방향의 변화 정보를 함께 사용하여 오디오 변곡점 검출을 더 정교하게 검출하게 되어 특정 화자의 발화구간 검출 성능 개선과 더 높은 품질의 음향신호 취득의 이점이 있다.

[0014] 본 발명의 바람직한 일 실시예에서는 또한, 호출어 발화 구간을 먼저 검출한 후에 임의의 묵음길이 후에 따라오는 질의어 발화구간 검출 과정에서, 호출어 발화 구간과 질어어 발화구간과의 화자 유사도를 측정하여 같은 화자가 발화한 질의어 발화구간을 더 정교하게 검출할 수 있는 효과가 있다.

[0015] 본 발명의 바람직한 일 실시예에서는 화자인식 기능이 구비된 조건 하에서 상기 오디오 변곡점 검출 기능과 화자식별 기능을 결합하여 특정 화자의 발화구간 검출과 사용자 식별 성능을 높임으로써 스마트 기기에서의 음성 인터페이스 서비스 및 어플리케이션의 품질을 올릴 수 있는 이점이 있다.

도면의 간단한 설명

[0016] 도 1 은 본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 시스템을 도시한다.

도 2 는 본 발명의 바람직한 일 실시예로서, 화자의 발화구간을 검출하는 전자장치의 내부 구성도를 도시한다.

도 3 은 본 발명의 바람직한 일 실시예로서, 오디오 변곡점을 화자의 발화구간 검출에 이용하는 일 실시예를 도시한다.

도 4 내지 5 는 본 발명의 바람직한 일 실시예로서, 오디오 변곡점 및 음원방향 정보를 화자의 발화구간 검출에 이용하는 일 실시예를 도시한다.

도 6 은 본 발명의 바람직한 일 실시예로서, 오디오 변곡점을 이용하여 화자의 발화구간 검출시 발화구간을 미세조정하는 일 예를 도시한다.

도 7 은 본 발명의 바람직한 일 실시예로서, 오디오 변곡점을 이용하여 질의어 발화구간 후보를 검출하기 위한 과정의 일 예를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0017] 이하, 본 발명의 다양한 실시 예가 첨부된 도면을 참조하여 기재된다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 실시 예의 다양한 변경(modification), 균등물(equivalent), 및/또는 대체물(alternative)을 포함하는 것으로 이해되어야 한다. 도면의 설명과 관련하여, 유사한 구성요소에 대

해서는 유사한 참조 부호가 사용될 수 있다.

[0018] 본 문서의 다양한 실시 예들에 따른 전자 장치는, 예를 들면, 스마트폰(smartphone), 태블릿 PC(tablet personal computer), 이동 전화기(mobile phone), 영상 전화기, 전자책 리더기(e-book reader), 데스크탑 PC(desktop PC), 랩탑 PC(laptop PC), 넷북 컴퓨터(netbook computer), 워크스테이션(workstation), 서버, PDA(personal digital assistant), PMP(portable multimedia player), MP3 플레이어, 모바일 의료기기, 카메라, 또는 웨어러블 장치(wearable device) 중 적어도 하나를 포함할 수 있다. 다양한 실시 예에 따르면 웨어러블 장치는 액세서리 형(예: 시계, 반지, 팔찌, 발찌, 목걸이, 안경, 콘택트 렌즈, 또는 머리 착용형 장치(head-mounted-device(HMD)), 직물 또는 의류 일체 형(예: 전자 의복), 신체 부착 형(예: 스킨 패드(skin pad) 또는 문신), 또는 생체 이식 형(예: implantable circuit) 중 적어도 하나를 포함할 수 있다.

[0019] 본 발명의 또 다른 일 실시예에서, 전자 장치는 가전 제품(home appliance)일 수 있다. 가전 제품은, 예를 들면, 텔레비전, DVD 플레이어(Digital Video Disk player), 오디오, 냉장고, 에어컨, 청소기, 오븐, 전자레인지, 세탁기, 공기 청정기, 셋톱 박스(set-top box), 홈 오토매이션 컨트롤 패널(home automation control panel), 보안 컨트롤 패널(security control panel), TV 박스(예: 삼성 HomeSync™, 애플TV™, 또는 구글 TV™), 게임 콘솔(예: Xbox™, PlayStation™), 전자 사전, 전자 키, 캠코더, 또는 전자 액자 중 적어도 하나를 포함할 수 있다. 이 외에 음성인터페이스를 통해 음성인식 및 음성합성을 수행하는 단말기를 모두 포함한다.

[0020] 도 1 은 본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 시스템을 도시한다.

[0021] 다채널오디오스트리밍입력에서 화자의 발화구간을 검출하는 시스템(100)은 복수의 마이크로폰(111, 112, 113, 114)을 포함하는 마이크로폰 어레이(110), 음성인식 내지 음성합성을 수행하는 전자장치(120), 오디오를 생성하는 적어도 하나의 단말기(130,131)를 포함한다. 마이크로폰 어레이(110)는 복수의 마이크로폰(111, 112, 113, 114)을 선형 또는 원형 구조로 배치하여 입력 음향을 녹음하도록 구현된다.

[0022] 본 발명의 바람직한 일 실시예로서, 전자장치(120)는 주변환경에서 다채널오디오스트리밍입력을 수신한다. 전자장치(120)는 하나 또는 복수의 마이크로폰(111, 112, 113, 114)으부터의 발성,발언 외에 다양한 단말기(130, 131)로부터의 소음, 전자장치(120)가 배치된 주변환경에서 발생되는 불필요한 음성, 소리, 소음을 포함할 수 있다. 이 경우, 음성인식이 필요한 특정 화자의 발화와 적어도 하나의 단말기(130,131)의 소리, 음성, 전자장치(120)가 배치된 주변환경에서 획득되는 다양한 노이즈, 소리, 음성 등을 동시에 수신할 수 있다.

[0023] 본 발명의 또 다른 바람직한 일 실시예로서, 마이크로폰 어레이(110)는 빔포머(Beam Former)(115)를 더 포함하여 음성인식을 수행할 수 있다. 빔포머(115)는 주변환경에서 수신되는 다채널오디오스트리밍입력의 각 채널이 특정 방향으로 음성을 분리하도록 다채널의 음성을 출력할 수 있다. 전자장치(120)는 빔포머(115)로부터 다채널오디오스트리밍입력을 수신하도록 구현될 수 있다. 전자장치(120)는 빔포머(115)를 통해 또는 하나 또는 복수의 마이크로폰(111, 112, 113, 114)으부터 수신한 다채널오디오스트리밍을 입력하여 특정화자의 호출어를 인식하고, 동일한 특정화자의 질의어를 식별하도록 구현된다.

[0024] 도 2 는 본 발명의 바람직한 일 실시예로서, 음성인식을 수행하는 전자장치의 내부 구성도를 도시한다.

[0025] 음성인식을 수행하는 전자장치(200)는 수신부(210), 제어부(220), 메모리(230) 및 출력부(240)를 포함한다. 또한 데이터 송수신을 위한 통신부와 디스플레이, 센서부 등 음성인식 및 음성합성에 요구되는 구성을 더 포함할 수 있다.

[0026] 제어부(220)는 다채널오디오스트리밍처리부(222), 오디오변곡점 검출부(224), 호출어발화구간 검출부(226), 질의어발화구간후보 검출부(250) 및 질의어발화구간 판정부(260)를 포함한다. 호출어발화구간 검출부(226)는 호출어를 인식하는 호출어 인식부(228)를 더 포함할 수 있다. 이 외에도 제어부(220)는 화자인식을 수행하는 화자인식부(미 도시), 질의어발화구간에서 인식한 화자의 질의어에 대응하는 음성을 합성한 음성음답 또는 텍스트 응답을 생성하는 응답생성부(미 도시)를 더 포함할 수 있다.

[0027] 출력부(240)는 특정 화자의 질의어 발화를 인식하여 자연어 처리, 응답생성, 음성합성 과정을 통해 생성된 오디오를 출력하거나, 음악 재생을 수행한다. 메모리(230)는 호출어에 대한 음향모델, 등록화자 모델, 사용자 프로파일 정보 등을 저장한다.

[0028] 전자장치(200)는 수신부(210)를 통해 다채널오디오스트리밍을 수신하여, 다채널오디오스트리밍처리부(222)에서 음성처리에 요구되는 전처리를 수행한다. 다채널오디오스트리밍처리부(222)는 예를 들어 반향제거, 음원방향 추

정, 빔포밍(beamforming), 음원분리 등 다양한 기능들을 이용하여 다채널의 입력 오디오스트리밍을 처리할 수 있다.

[0029] 본 발명의 바람직한 일 실시예로서, 다채널오디오스트리밍처리부(222)는 복수 개의 채널오디오스트리밍이 동시에 또는 상이하게 입력될 때, 마이크로폰 채널 쌍간에 도착시간지연 특징을 추출하여 음원방향 정보를 추출할 수 있다. 이 경우 다채널오디오스트리밍처리부(222)는 일 예를 들어 LMS(least mean square) 적응 필터를 이용하여 마이크로폰 채널 쌍간에 도착시간지연 특징들을 계산하고, N개의 가장 가능성 있는 도착시간지연 특징을 추출하여 음원방향 정보를 추출할 수 있다. 또한, 다채널오디오스트리밍처리부(222)는 GCC-PHAT(generalized cross correlation phase transform) 기법에 기반한 지연 및 합계(delay and sum)알고리즘을 적용하여 유효한 수의 도착시간지연 특징을 추출하여 음원방향 정보를 추출할 수 있다. 이에 대해서는 도 4에서 보다 상세히 살펴본다.

[0030] 오디오변곡점 검출부(224)는 다채널오디오스트리밍처리부(222)에서 음성처리에 요구되는 전처리가 수행된 다채널오디오스트리밍입력에서 오디오 변곡점을 적어도 하나 이상 추출한다. 오디오변곡점 검출은 입력되는 오디오스트리밍에서 앞뒤 오디오 특성이 달라지는 지점을 찾는 기술로, 오디오 변곡점은 음악에서 음성으로, 잡음에서 음성으로, 또는 묵음에서 음성으로 등과 같이 음원 속성이 변경되거나 또는 발화하는 화자가 변경되는 지점을 나타낸다. 오디오 변곡점을 검출하는 방법은 거리척도 변화 그래프에서 로컬 최대값(local maximum) 탐색등에 의해 피크(peak)값을 검출하는 방법을 이용할 수 있다. 예를 들어, BIC(Bayesian Information Criterion), Generalized Likelihood Ratio 등의 다양한 척도를 이용할 수 있다. 기존에는 오디오 변곡점을 검출하기 위해 일정한 시간간격 단위로, 예를 들어 10msec, MFCC와 같은 특징벡터 추출한 후 일정 주기별로, 예를 들어 100msec, 좌우 세그먼트 사이의 거리를 계산하는 방법을 이용한다. 좌우 세그먼트는 각각 1초에서 3초 사이의 길이일 수 있다.

[0031] 이와 달리 본 발명의 바람직한 일 실시예에서는, 오디오변곡점 검출부(224)는 피치(pitch)정보를 기반으로 오디오 변곡점을 검출한다. 도 3을 더 참고하면 본 발명의 바람직한 일 실시예에서는 오디오변곡점 검출부(224)는 입력되는 오디오스트리밍에서 특징벡터를 추출한 후 안정적인 피치 지속길이(도 3, 330)를 계산하여 피치의 변곡점을 검출한다. 이 경우 안정적인 피치 지속길이는 다채널오디오스트리밍 신호를 전처리 수행하여 일정 길이마다 음성 특징벡터를 추출한 후 기계학습을 통해 안정적이라고 학습되는 피치 지속길이를 산출할 수 있다.

[0032] 일 실시예로서, 30msec의 음성 입력신호에 대해 10msec 의 프레임 단위로 이동하면서 음성 분석을 진행하는 경우, 본 발명의 바람직한 일 실시예에서는 매 프레임마다 주요 주파수 성분의 주기를 나타내는 피치를 구하고, 기설정된 유사성을 만족하는 피치가 지속되는 영역을 안정적인 피치 지속길이라고 지칭한다. 매 프레임마다 피치를 구하고, 이전 피치와의 유사성을 만족하면 안정적인 피치구간이 한 프레임씩 증가시키는 형태로 기계학습이 가능하다. 일반적으로 동일 화자의 모음 구간에서는 동일한 피치들이 연속적으로 나타나는 특성을 보이므로 안정적인 피치구간은 동일화자의 특정 모음에 대한 발성 구간으로 판정할 수 있다.

[0033] 그 후, 피치의 변곡점(도 3, 340)을 기준으로 좌우 세그먼트간 거리를 계산한다. 계산된 거리를 기초로 생성한 거리척도 그래프(도 3, 350)에서 먼저 거리 척도 값이 기설정된 기준치 이상이 되는 지점(도 3, 350a)들을 먼저 선택하고, 일정 구간 내에 복수 개가 존재할 경우 최대 값을 갖는 지점만 변곡점으로 선택한다(도 3, 351, 352, 353, 354, 355, 356). 본 발명의 바람직한 일 실시예에서는 피치 변곡점에 대해서만 오디오 변곡점을 검출함으로써 기존의 방법에 비해 성능이 개선되고 계산량이 감축되는 효과가 있다.

[0034] 본 발명의 또 다른 바람직한 일 실시예에서, 오디오변곡점 검출부(224)는 피치(pitch)정보 외에 추가로 다채널오디오스트리밍처리부(222)에서 도착시간지연 특징을 추출하여 파악한 음원방향 정보를 더 이용하여 오디오 변곡점을 검출할 수 있다. 이에 대해서는 도 4 내지 5를 더 참고하여 설명한다.

[0035] 다채널오디오스트리밍처리부(222)에서 제 1 채널의 오디오스트리밍 신호(410)를 수신하고, 빔포밍 출력 오디오스트리밍신호(420)를 이용하여 음성 특징을 추출하고, 안정적인 피치 지속길이(430)를 계산한다. 그 후, 피치 변곡점을 검출(440)하고, 피치의 변곡점을 기준으로 좌우 세그먼트간 거리를 계산할 때 도착시간지연 특징벡터를 함께 이용한다. 이를 위해 일 실시예로서, GCC-PHAT 알고리즘 이용하여 유효한 수의 도착시간지연 특징을 추출함으로써 음원의 위치를 추적할 수 있다(도 4, 451, 452). 이 후, 추적한 음원의 방향(도 4, 451, 452) 및 피치 변곡점 정보(도 4, 440)를 모두 이용하여 좌우 세그먼트간 거리를 계산한다. 그 후 계산된 거리를 기초로 생성한 거리척도 그래프(도 4, 460)에서 피크 정보를 이용하여 변곡점(도 4, 461, 462, 463, 464)을 검출한다.

[0036] 도 5 는 본 발명의 바람직한 일 실시예로서, 오디오변곡점 검출부(224)는 피치(pitch)정보 및 도착시간지연 특

징을 모두 이용하여 오디오 변곡점을 검출하는 일 실시예를 도시한다. 다채널오디오스트리밍처리부(222)는 수신한 다채널오디오스트리밍 $x_1[n]$(520a), $x_2[n]$(520b),..,$x_N[n]$(520c) 각각에 대해 잡음을 제거한 후 $x_1'[n]$(521a), $x_2'[n]$(521b),..,$x_N'[n]$(521c), 다채널간 상호상관 정도를 계산하고(S510), 도착시간지연 특징을 추출한다(S520). 이 후, 빔포밍 등과 같은 전처리(S530)를 수행할 수 있다. 빔포밍 기법을 적용하여 전처리(S530)를 수행하는 경우 오디오 변곡점 이전의 방향들에 대한 신호를 상쇄시켜 개선된 품질의 음성을 추출할 수 있다.

[0037] 다채널오디오스트리밍처리부(222)는 다채널오디오스트리밍에 대해 전처리를 수행한 빔포밍오디오스트리밍 신호 y[n]을 오디오변곡점검출부(224)로 전송한다. 오디오변곡점검출부(224)는 수신한 신호에서 특징벡터를 추출(S540)한 후, 추출한 특징벡터를 이용하여 피치정보를 추출하여 안정적인 피치 지속길이(stable pitch duration)를 결정하고, 피치변곡점을 검출한다(S550). 그 후, 오디오변곡점검출부(224)는 피치변곡점 및 다채널오디오스트리밍처리부(222)에서 추출한 도착시간지연 특징정보를 이용하여 좌우 세그먼트간 거리를 계산하고(S560), 계산된 거리로 생성한 거리척도 그래프에서 피크(peak) 추정으로 오디오 변곡점을 검출한다(S570).

[0038] 도 2 로 돌아와서, 호출어발화구간 검출부(226)는 다채널오디오스트리밍입력 중 검출된 적어도 하나의 오디오 변곡점 지점들만을 대상으로 호출어 모델과의 유사도를 측정하여 화자의 호출어발화구간을 검출한다. 기존에는 모든 순간을 호출어의 시작 또는 끝이 가능한 순간으로 판단하여 계산을 수행하였으나, 본 발명의 바람직한 일 실시예에서는 오디오 변곡점 지점들만을 대상으로 기지정된 호출어 모델과의 유사도를 측정함으로써 계산량을 줄일 수 있는 효과가 있다.

[0039] 호출어발화구간 검출부(226)는 HMM(hidden Markov model) 알고리즘을 이용하여 호출어 발화구간의 음성과 지정된 호출어 모델과의 매칭 여부를 검사한다. 이 경우 음성신호를 상태 천이 확률과 각 상태에서의 관찰확률이라는 두 단계의 확률 과정으로 표현하며, 관측확률은 GMM(Gaussian mixture model), 또는 DNN(deep neural network)에 의해 모델링될 수 있다. 도 6을 참고하면, 호출어발화구간 검출부(226)에서 오디오 변곡점 지점(630, 631, 632, 633, 634, 635)을 기준으로 호출어 모델과의 유사도를 HMM 알고리즘으로 측정하여 화자의 호출어발화구간(S630)을 검출하는 일 실시예를 도시한다.

[0040] 호출어발화구간의 시작점(610)과 끝(620)은 오디오 변곡점 중에서 선택된다. 본 발명의 바람직한 일 실시예에서는 오디오 변곡점 검출에서의 정교한 지점 감지에서 발생할 수 있는 미세한 오류를 보상하기 위해, 변곡점 기준으로 앞뒤 일정 길이의 범위에서 호출어 시작과 끝의 가능하도록 허용할 수 있다. 예를 들어, 시작점(610)과 끝(620) 지점을 기초로 (-50, +50) 밀리세컨드 범위(S610, S620)에서 호출어의 시작 및 끝이 가능하도록 허용할 수 있다.

[0041] 도 5에서 x축은 다채널오디오스트리밍입력(510)의 특징벡터 열을 나타내고, y축은 HMM 알고리즘으로 측정한 호출어 모델과의 유사도를 나타낸다. HMM 알고리즘과 관련된 상세한 내용은 본 발명이 속하는 통상의 지식을 가진 자에게 자명한 바 상세한 설명을 생략한다.

[0042] 본 발명의 또 다른 바람직한 일 실시예로서, 메모리(230)에 복수의 등록된 화자성문모델이 있는 경우, 호출어발화구간 검출부(226)에서 검출한 발화구간에 대해 추가적으로 화자인식 또는 화자검증을 수행할 수 있다. 화자인식 수행과정에서 화자성문모델과의 매칭점수가 기준치 이상인 경우 등록된 화자로, 미만인 경우 미등록화자로 판정한다. 화자성문모델은 GMM, SVM(support vector machine), i-vector 등을 이용하여 기계학습이 가능하다. 호출어발화구간검출부(226)에서 호출어 발화구간을 검출되면, 호출어 인식부(228)는 호출어 발화구간 내에서 기지정된 호출어를 인식할 수 있다.

[0043] 질의어발화구간후보 검출부(250)는 다채널오디오스트리밍입력에서 호출어발화구간 검출부(226)에서 검출된 호출어 발화구간 이후에 뒤따라오는 구간에서 검출된 적어도 하나의 오디오 변곡점의 조합을 기초로 적어도 하나의 질의어발화구간후보를 모두 검출한다. 도 7을 참고하여 설명한다.

[0044] 도 7 을 참고하면 다채널오디오스트리밍입력에서 검출된 호출어 발화구간 이후에 뒤따라오는 구간에서 오디오 변곡점이 검출된 지점(t1, t2, t3, t4, t5)들로 생성될 수 있는 모든 구간(S710, S711, S712, S713, S720, S721, S722, S730, S731, S740)을 질의어발화구간후보로 검출한다.

[0045] 그리고, 질의어발화구간 판정부(260)는 호출어발화구간 검출부(226)에서 검출된 화자와 적어도 하나의 질의어발화구간후보(S710, S711, S712, S713, S720, S721, S722, S730, S731, S740) 의 화자를 비교하여 화자유사도를 측정한 후, 화자유사도가 기설정된 값을 초과하는 구간들만을 선택적으로 검출한다.

[0046] 이 경우, 질의어발화구간 판정부(260)는 질의어 발화 시작이 가능한 시점, t1, t2, t3, t4 에서 시작되는 첫 구간(S710, S720, S730, S740)에서 검출된 화자의 음성이 호출어 발화구간에서 검출된 화자와의 화자유사도가 기설정된 값에 해당하지 않는 경우, 다음 구간에 대해서는 화자유사도를 측정하지 않는다.

[0047] 일 예를 들어, t1 시점을 질의어 시작 지점으로 예측하여 t1과 t2 구간(S710)에서 검출한 화자와 호출어 발화구간에서 검출된 화자가 유사하지 않다고 판단되는 경우, t1과 t3 구간(S711), t1과 t4 구간(S712),t1과 t5 구간(S713)은 더 이상 화자유사도를 측정하지 않고, t2 시점을 질의어 시작 지점을 다시 예측하여 질의어발화구간후보를 검출할 수 있다. 이상의 방식으로, 질의어발화구간후보로 t2에서 t3구간(S720), t2에서 t4구간(S721), 그리고, t3에서 t4구간(S731)이 검출될 수 있다.

[0048] 질의어발화구간 판정부(260)는 선택된 t2에서 t3구간(S720), t2에서 t4구간(S721), 그리고, t3에서 t4구간(S731) 질의어발화구간 후보들 중 최장의 발화구간, 예를 들어 t2에서 t4구간(S721),을 질의어발화구간으로 판정한다.

[0049] 본 발명의 바람직한 일 실시예로서, 질의어발화구간 판정부(260)는 화자유사도 측정시 오디오 변곡점 검출부(224)에서 오디오 변곡점 검출시 사용한 BIC와 같은 거리 척도 알고리즘을 사용할 수 있다. 또한, 호출어 발화구간의 i-vector와 특정 조합의 질의어 발화의 i-vector 사이의 유사도를 나타내는 매칭점수에 기초하여 유사도를 판정할 수 있다. 이 외에도 다양한 방법을 이용할 수 있다.

[0050] 본 발명의 바람직한 일 실시예로서, 질의어발화구간 판정부(260)는 호출어발화구간 검출부(226)에서 인식된 호출어가 등록된 화자의 발화로 식별된 경우, 질의어발화구간 판정부(260)는 식별된 등록화자의 텍스트독립형 성문모델과의 식별점수를 기준으로 특정화자의 발화인지 여부를 판단할 수 있다. 또한, 호출어 발화구간과의 화자 유사성 및 식별화자와의 유사성 모두를 이용할 수 있다.

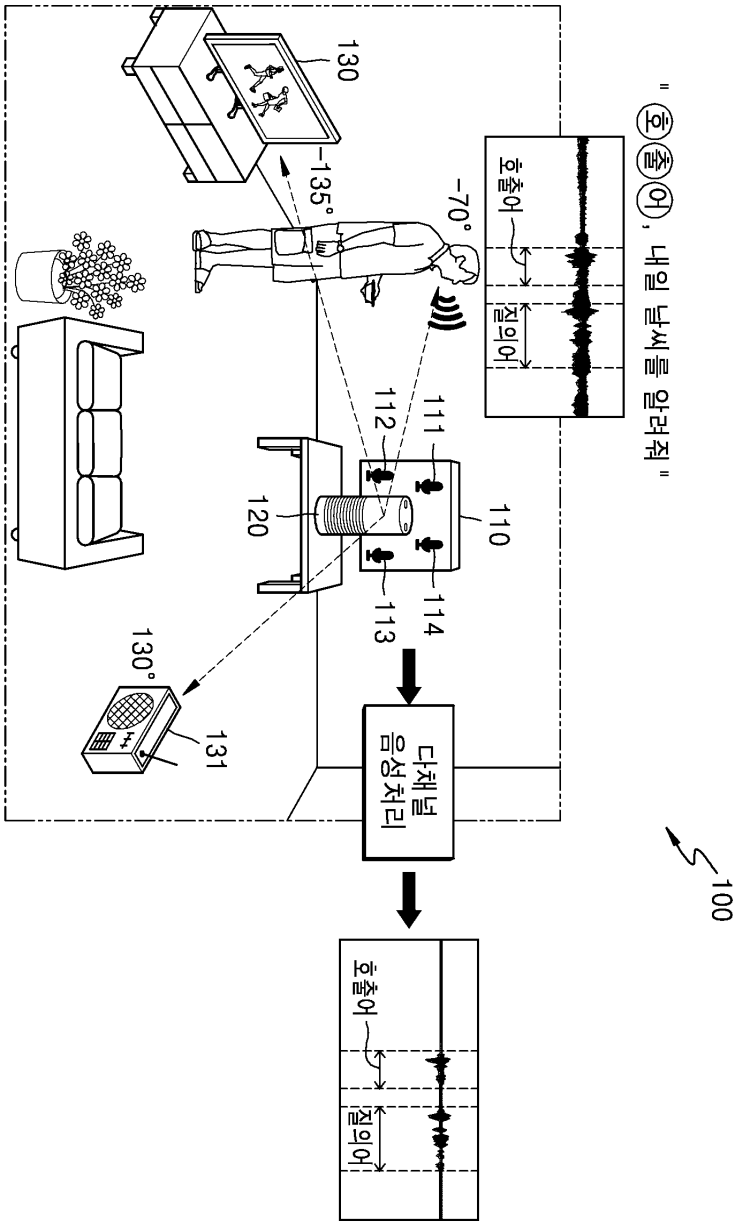
[0051] 이상의 방식을 통해 호출어발화구간 검출부(226)와 질의어발화구간 판정부(260) 각각에서 동일한 화자를 식별하고, 동일한 화자가 발화한 호출어와 질의어를 각각 식별할 수 있다. 그 후, 제어부(220)에서 식별한 질의어에 대응하는 액션(action) 또는 응답을 출력부(240)를 통해 출력할 수 있다.

[0052] 본 발명의 실시예들은 다양한 컴퓨터로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터 판독 가능 매체를 포함한다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프린터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

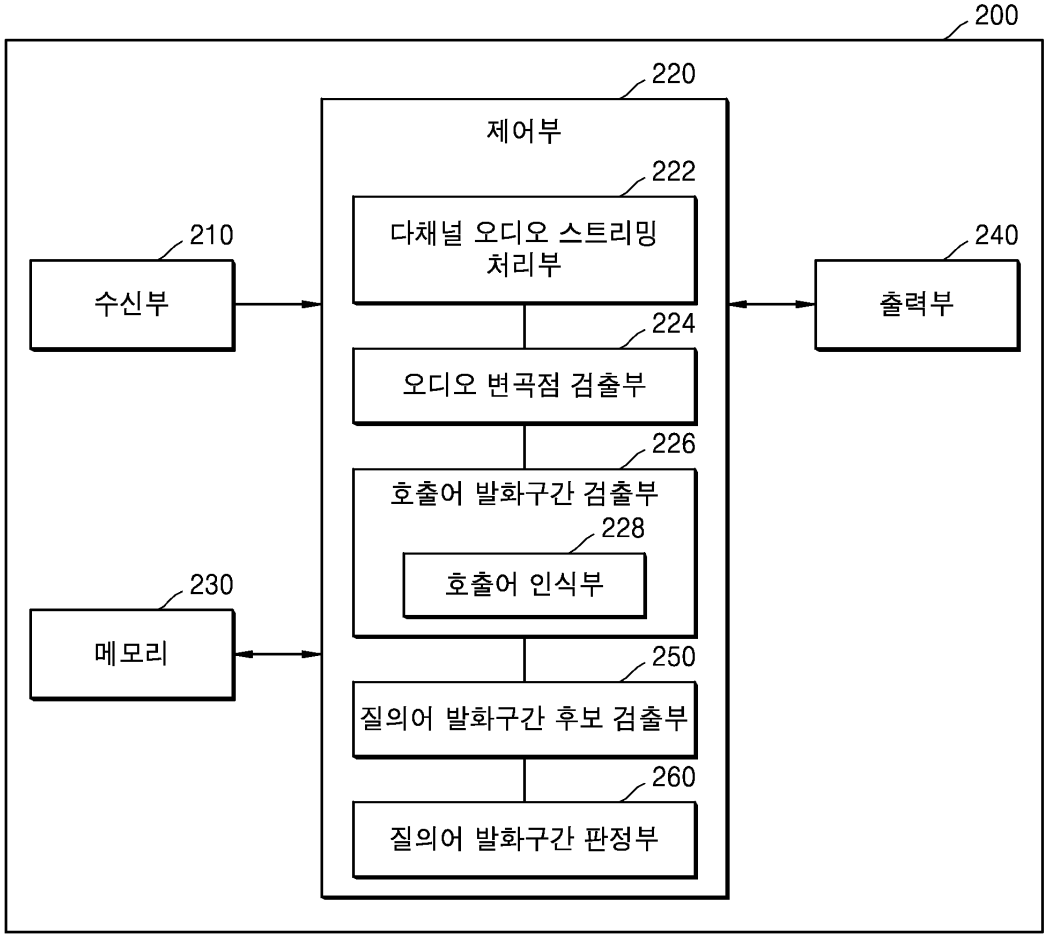
[0053] 이상에서는 본 발명의 바람직한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특정의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야 에서 통상의 지식을 가진자에 의해 다양한 변형실시가 가능한 것은 물론이고, 이러한 변형실시들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어서는 안될 것이다.

도면

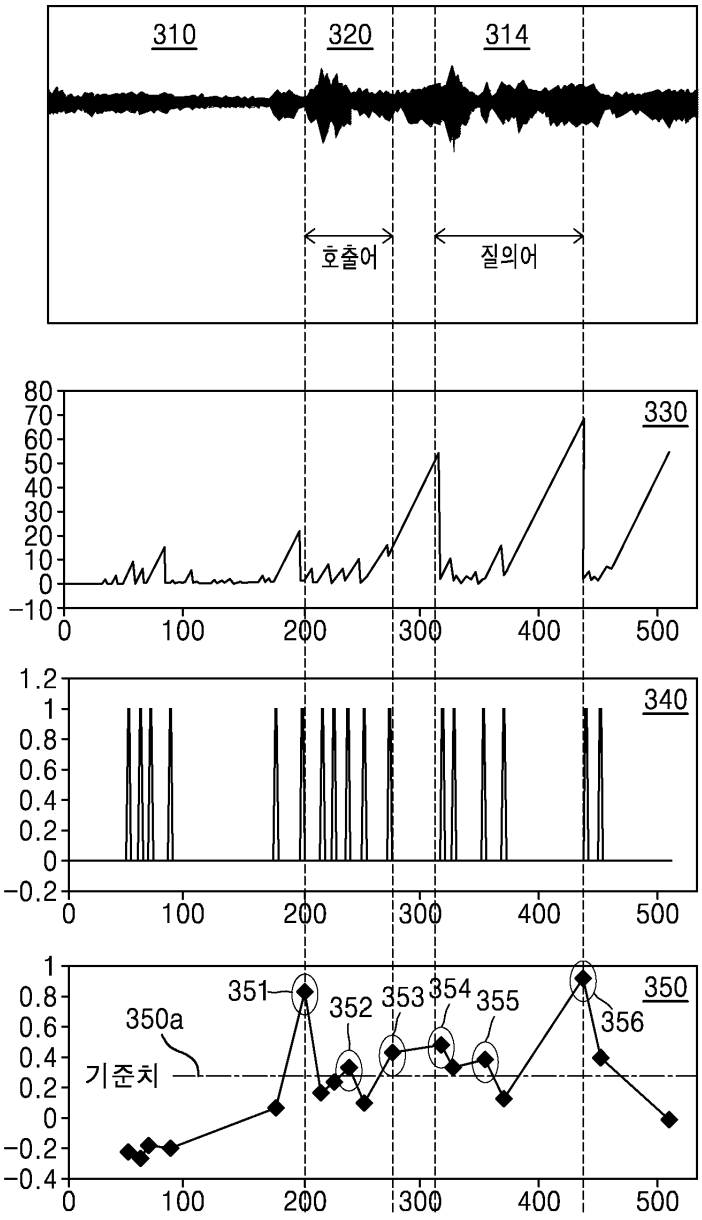
도면1


"호출어, 내일 날씨를 알려줘"
호출어
질의어
130
-135°
-70°
112
111
110
120
113
114
130°
131
다채널 음성처리
100
호출어
질의어

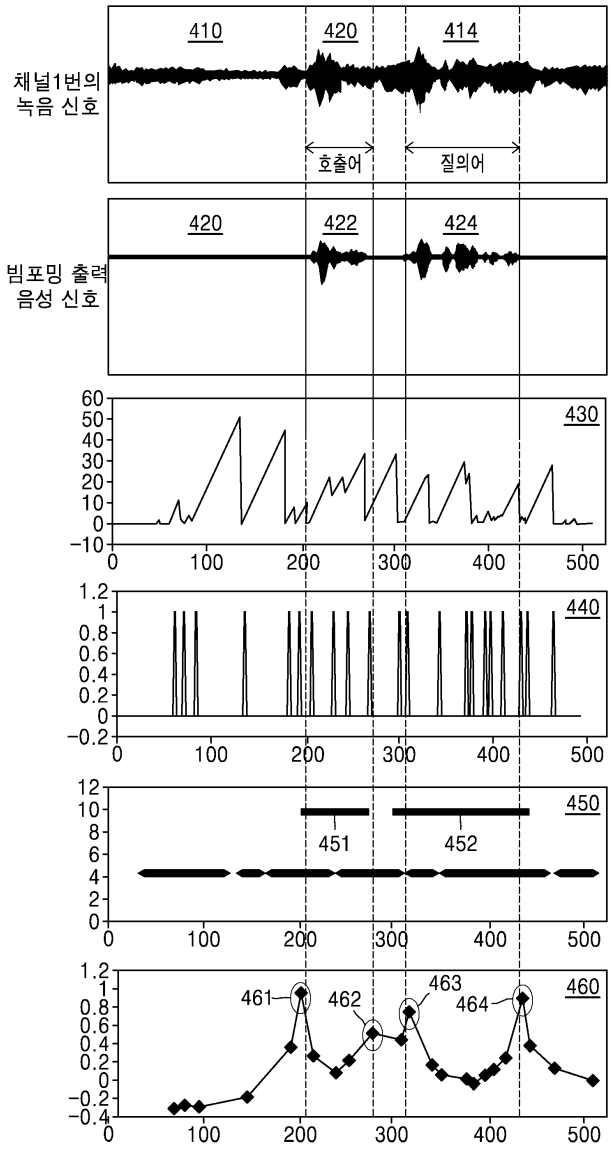
도면2


200
220
제어부
222
다채널 오디오 스트리밍 처리부
210
수신부
240
출력부
224
오디오 변곡점 검출부
226
호출어 발화구간 검출부
228
호출어 인식부
230
메모리
250
질의어 발화구간 후보 검출부
260
질의어 발화구간 판정부

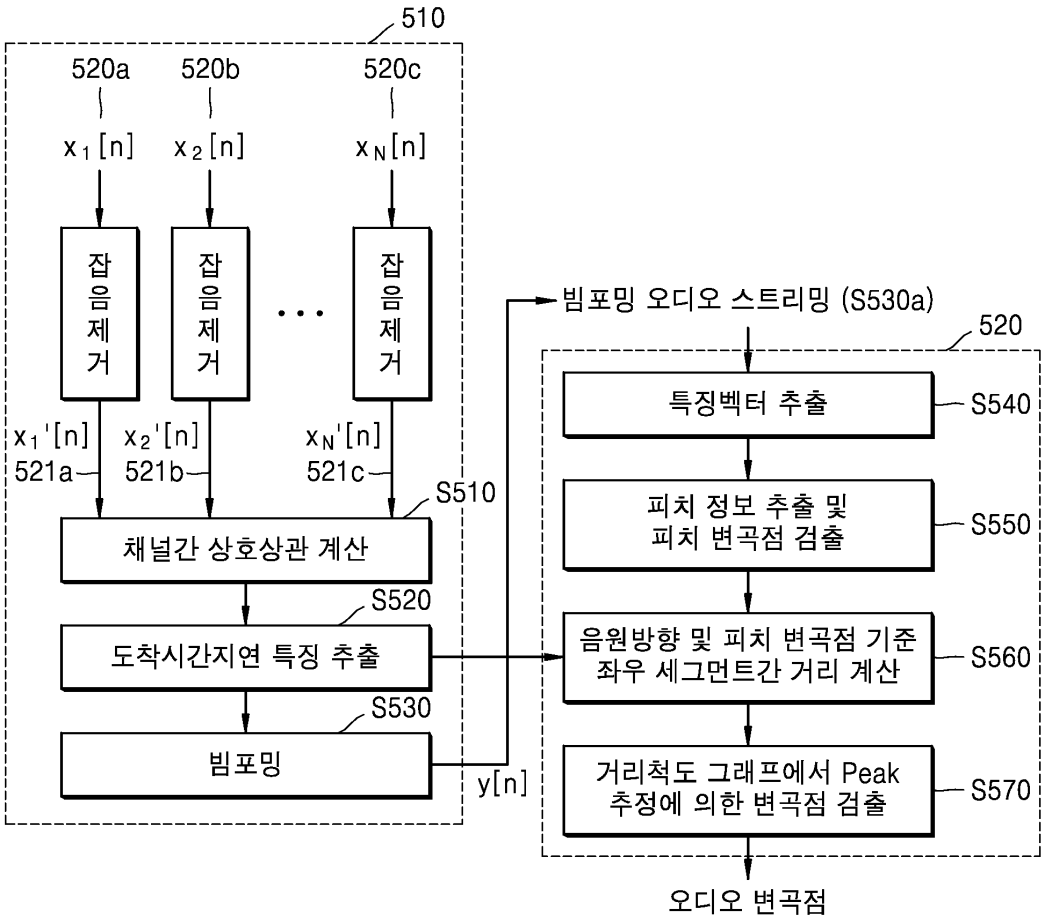
도면3



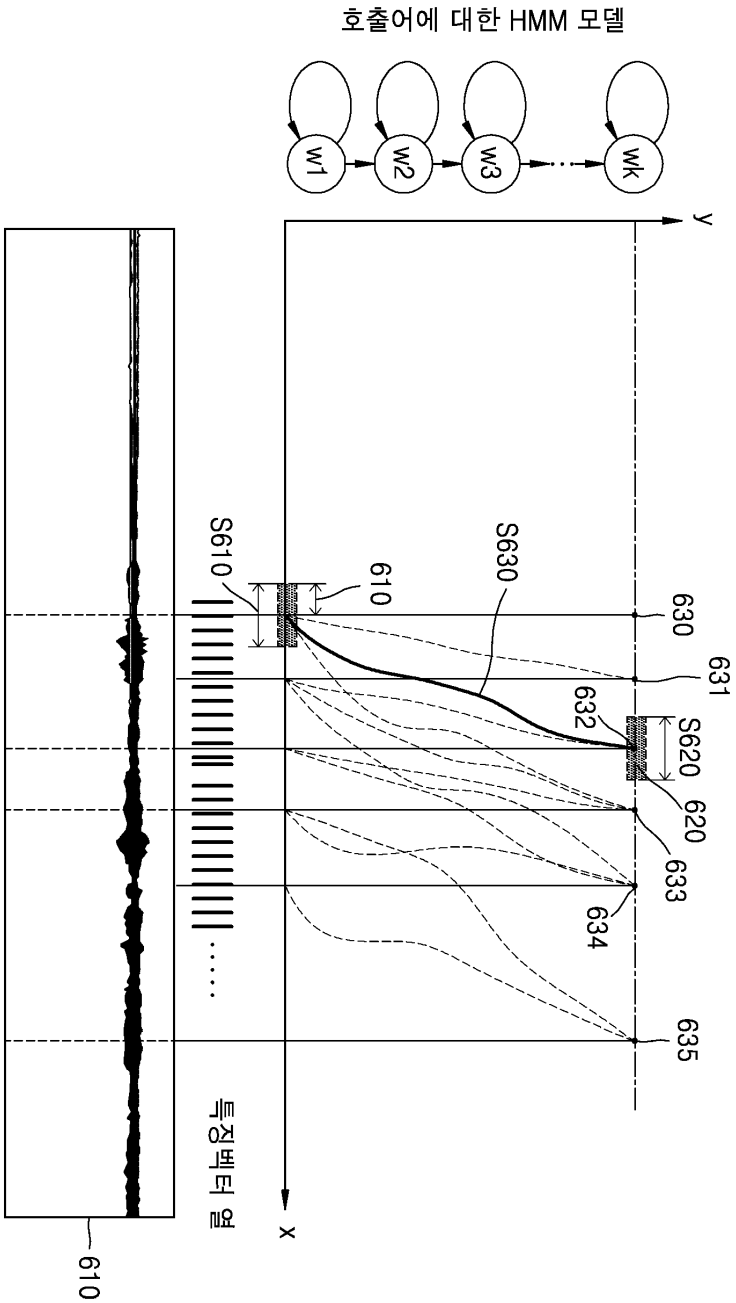
도면4


410
420
414
채널1번의
녹음 신호
호출어
질의어
420
422
424
빔포밍 출력
음성 신호
430
440
450
451
452
460
461
462
463
464

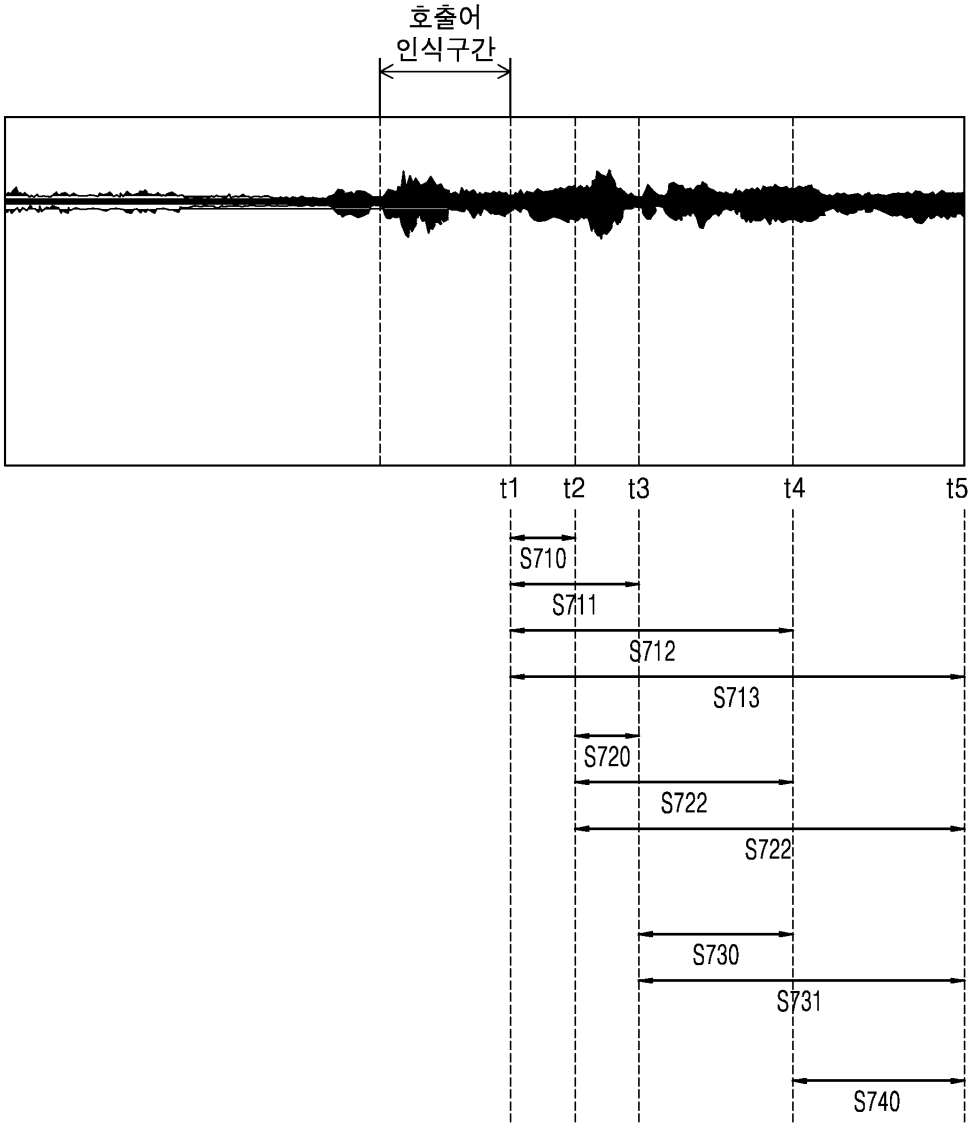
도면5


510
520a
520b
520c
x1[n]
x2[n]
xN[n]
잡음제거
잡음제거
잡음제거
x1'[n]
521a
x2'[n]
521b
xN'[n]
521c
S510
채널간 상호상관 계산
S520
도착시간지연 특징 추출
S530
빔포밍
y[n]
빔포밍 오디오 스트리밍 (S530a)
520
특징벡터 추출
S540
피치 정보 추출 및
피치 변곡점 검출
S550
음원방향 및 피치 변곡점 기준
좌우 세그먼트간 거리 계산
S560
거리척도 그래프에서 Peak
추정에 의한 변곡점 검출
S570
오디오 변곡점

도면6


호출어에 대한 HMM 모델
w1
w2
w3
wk
y
x
S610
610
S630
630
631
632
S620
620
633
634
635
특징벡터 열
610

도면7


호출어
인식구간
t1
t2
t3
t4
t5
S710
S711
S712
S713
S720
S722
S722
S730
S731
S740