(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0144233
(43) 공개일자 2014년12월18일

(51) 국제특허분류(Int. Cl.)
*G10L 15/02* (2006.01) *G10L 15/07* (2013.01)
(21) 출원번호 10-2014-7029482
(22) 출원일자(국제) 2013년07월08일
심사청구일자 2014년10월21일
(85) 번역문제출일자 2014년10월21일
(86) 국제출원번호 PCT/CN2013/079005
(87) 국제공개번호 WO 2014/008843
국제공개일자 2014년01월16일
(30) 우선권주장
201210235593.0 2012년07월09일 중국(CN)

(71) 출원인
**후아웨이 디바이스 컴퍼니 리미티드**
중국 션젠 롱강 디스트릭트 반티안 후아웨이 인더스트리얼 베이스 빌딩 비2
(72) 발명자
**루 팅**
중국 518129 광동 션젠 롱강 반티안 후아웨이 어드미니스트레이션 빌딩
(74) 대리인
**유미특허법인**

전체 청구항 수 : 총 10 항

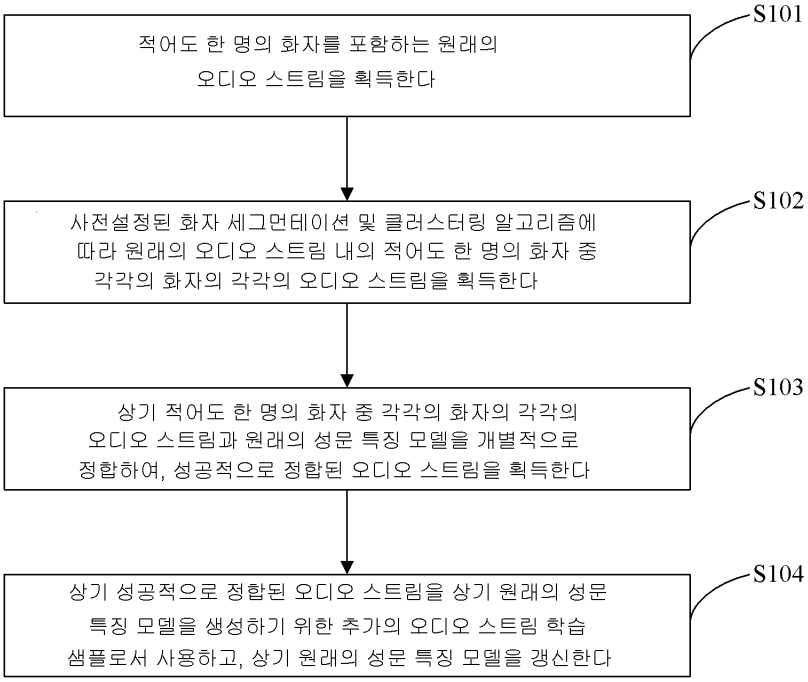
(54) 발명의 명칭 **성문 특징 모델 갱신 방법 및 단말**

## (57) 요 약

본 발명은 음성 인식 기술 분야에 적용 가능하며, 성문 특징 모델 갱신 방법 및 단말을 제공한다. 성문 특징 모델 갱신 방법은: 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하는 단계; 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자

(뒷면에 계속)

*대 표 도* - 도1



의 각각의 오디오 스트림을 획득하는 단계; 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하는 단계; 및 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신하는 단계를 포함한다. 본 발명에서는, 호출 동안 유효한 오디오 스트림이 적응적으로 추출되고 추가의 오디오 스트림 학습 샘플로서 사용되어, 원래의 성문 특징 모듈을 동적으로 정정하며, 이에 의해 상대적으로 높은 실용성의 전제 하에 성문 특징 모델의 정확도 및 인식 정확도를 높이는 목적을 달성한다.

## *특허청구의 범위*

**청구항 1**

성문(voiceprint) 특징 모델 갱신 방법에 있어서,

적어도 한 명의 화자(speaker)를 포함하는 원래의 오디오 스트림을 획득하는 단계;

사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하는 단계;

상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하는 단계; 및

상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신하는 단계

를 포함하는 성문 특징 모델 갱신 방법.

**청구항 2**

제1항에 있어서,

상기 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하는 단계 이전에,

사전설정된 오디오 스트림 학습 샘플에 따라 원래의 성문 특징 모델을 확립하는 단계

를 더 포함하는 성문 특징 모델 갱신 방법.

**청구항 3**

제1항 또는 제2항에 있어서,

사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하는 단계는 구체적으로,

상기 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하는 단계 - 상기 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함함 - ; 및

사전설정된 화자 클러스터링 알고리즘에 따라, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성하는 단계

를 포함하는, 성문 특징 모델 갱신 방법.

**청구항 4**

제1항 내지 제3항 중 어느 한 항에 있어서,

상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하는 단계는,

상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 상기 원래의 성문 특징 모델에 따라, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도(matching degree)를 획득하는 단계; 및

가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 상기 성공적으로 정합된 오디오 스트림으로 선택하는 단계

를 포함하는, 성문 특징 모델 갱신 방법.

**청구항 5**

제1항 내지 제4항 중 어느 한 항에 있어서,

상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신하는 단계는 구체적으로,

상기 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하는 단계 - 상기 사전설정된 오디오 스트림 학습 샘플은 상기 원래의 성문 특징 모델을 생성하기 위한 오디오 스트림임 - ; 및

상기 정정된 성문 특징 모델에 따라 상기 원래의 성문 특징 모델을 갱신하는 단계

를 포함하는, 성문 특징 모델 갱신 방법.

**청구항 6**

단말에 있어서,

원래의 오디오 스트림 획득 유닛, 세그먼테이션 및 클러스터링 유닛, 정합 유닛, 및 모델 갱신 유닛

을 포함하며,

상기 원래의 오디오 스트림 획득 유닛은 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하고 상기 원래의 오디오 스트림을 상기 세그먼테이션 및 클러스터링 유닛에 송신하도록 구성되어 있으며;

상기 세그먼테이션 및 클러스터링 유닛은 상기 원래의 오디오 스트림 획득 유닛에 의해 송신된 원래의 오디오 스트림을 수신하고, 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하며, 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 상기 정합 유닛에 송신하도록 구성되어 있으며;

상기 정합 유닛은 상기 세그먼테이션 및 클러스터링 유닛에 의해 송신된 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 수신하고, 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하며, 상기 성공적으로 정합된 오디오 스트림을 상기 모델 갱신 유닛에 송신하도록 구성되어 있으며; 그리고

상기 모델 갱신 유닛은 상기 모델 갱신 유닛에 송신된 상기 성공적으로 정합된 오디오 스트림을 수신하고, 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하며, 상기 원래의 성문 특징 모델을 갱신하도록 구성되어 있는, 단말.

**청구항 7**

제6항에 있어서,

샘플 획득 유닛 및 원래의 모델 확립 유닛을 더 포함하며,

상기 샘플 획득 유닛은 사전설정된 오디오 스트림 학습 샘플을 획득하고, 상기 사전설정된 오디오 스트림 학습 샘플을 상기 원래의 모델 확립 유닛에 송신하도록 구성되어 있으며,

상기 원래의 모델 확립 유닛은 상기 샘플 획득 유닛에 의해 송신된 상기 사전설정된 오디오 스트림 학습 샘플을 수신하고, 상기 사전설정된 오디오 스트림 학습 샘플에 따라 상기 원래의 성문 특징 모델을 확립하도록 구성되어 있는, 단말.

**청구항 8**

제6항 또는 제7항에 있어서,

상기 세그먼테이션 및 클러스터링 유닛은 구체적으로 세그먼테이션 유닛 및 클러스터링 유닛을 포함하며,

상기 세그먼테이션 유닛은 상기 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하며 - 상기 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함함 - 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을

상기 클러스터링 유닛에 송신하도록 구성되어 있으며; 그리고

상기 클러스터링 유닛은 상기 세그먼테이션 유닛에 의해 송신된, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 수신하고, 사전설정된 화자 클러스터링 알고리즘에 따라, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성하도록 구성되어 있는, 단말.

**청구항 9**

제6항 내지 제8항 중 어느 한 항에 있어서,

상기 정합 유닛은 구체적으로 정합도 획득 유닛 및 정합 오디오 스트림 획득 유닛을 포함하며,

상기 정합도 획득 유닛은 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 상기 원래의 성문 특징 모델에 따라, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 획득하고, 상기 정합도를 상기 정합 오디오 스트림 획득 유닛에 송신하도록 구성되어 있으며; 그리고

상기 정합 오디오 스트림 획득 유닛은 상기 정합도 획득 유닛에 의해 송신된, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 수신하고, 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 상기 성공적으로 정합된 오디오 스트림으로 선택하도록 구성되어 있는, 단말.

**청구항 10**

제6항 내지 제9항 중 어느 한 항에 있어서,

상기 모델 갱신 유닛은 구체적으로 정정 모델 획득 유닛 및 모델 갱신 서브유닛을 포함하며,

상기 정정 모델 획득 유닛은 상기 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하고, 상기 정정된 성문 특징 모델을 상기 모델 갱신 서브유닛에 송신하도록 구성되어 있으며; 그리고

상기 모델 갱신 서브유닛은 상기 정정 모델 획득 유닛에 의해 송신된 상기 정정된 성문 특징 모델을 수신하고, 상기 정정된 성문 특징 모델에 따라 상기 원래의 성문 특징 모델을 갱신하도록 구성되어 있는, 단말.

# 명 세 서

## 기 술 분 야

[0001] 본 출원은 2012년 7월 9일에 중국특허청에 출원되고 발명의 명칭이 "METHOD FOR UPDATING VOICEPRINT FEATURE MODEL AND TERMINAL"인 중국특허출원 No. 201210235593.0에 대한 우선권을 주장하는 바이며, 상기 문헌의 내용은 본 명세서에 원용되어 병합된다.

[0002] 본 발명은 음성 인식 기술 분야에 관한 것이며, 특히 성문 특징 모델 갱신 방법 및 단말에 관한 것이다.

## 배 경 기 술

[0003] 성문 인식은 사람의 소리를 사용함으로써 실행되는 인식 기술의 한 유형이다. 사람이 말을 할 때 사용되는 발성 기관 간에는 약간의 차이가 있으며, 임의의 두 사람 소리의 성문 스펙트로그램(voiceprint spectrogram)은 다르다. 그러므로 성문은 개인차를 나타내는 생물학적 특징으로 사용될 수 있다. 즉, 성문 특징 모델을 확립함으로써 서로 다른 개체를 나타낼 수 있는데, 이 성문 특징 모델을 사용하여 서로 다른 개체를 인식한다. 현재, 성문 특징 모델은 딜레마에 빠져 있는데, 이것은 학습 말뭉치의 길이 선택에 주로 반영되어 있다. 일반적으로, 성문 학습 말뭉치를 길게 하면 확립된 특징 모델을 더 정확해지고 인식 정확도가 높아지지만, 실용성이 떨어지고; 성문 학습 말뭉치를 짧게 하면 실용성은 좋아지지만, 인식 정확도가 높지 않다. 또한, 실제의 애플리케이션에서, 예를 들어, 화면 성문 잠금해제 애플리케이션에서, 보안을 충족하기 위해서는 높은 인식 정확도가 요구되고, 실용성을 좋게 하기 위해서는 학습 말뭉치가 과도하게 길어서는 안 된다.

## 발명의 내용

### 해결하려는 과제

[0004] 기존의 성문 특징 모델 확립 방법에서는, 사용자가 성문 등록 구문에서 학습을 복수 회 수행하고 각각의 학습에서 짧은 말뭉치를 사용하며, 최종적으로 이 짧은 말뭉치를 긴 학습 말뭉치와 결합하여 특징 모델을 생성한다. 그렇지만, 사용자는 특정한 기간 동안 복수 회 학습 말뭉치를 수동으로 기록할 때 안 좋은 경험을 할 수 있고; 그 결합된 학습 말뭉치의 길이는 여전히 제한적이고, 정확한 특징 모델이 생성될 수 없고, 인식 정확도는 더 향상될 수 없으며; 말하는 속도의 변화와 감정 동요 역시 모델 확립 정확도에 영향을 미칠 수 있다. 그러므로 성문 특징 모듈의 정확도를 어떻게 개선할 것인가와 상대적으로 높은 실용성의 전제 하에서 인식 정확도를 더 향상시키는 것이 시급한 과제이다.

### 과제의 해결 수단

[0005] 본 발명의 목적은 성문 특징 모델 갱신 방법 및 단말을 제공하여, 기존의 방법을 사용하여 성문 특징 모델을 획득할 때, 상대적으로 높은 실용성의 전제 하에서는 성문 특징 모델의 정확도를 개선하는 것이 확보될 수 없고, 그 결과 성문 특징 모델을 사용해서는 인식 정확도를 개선할 수 없다는 문제를 해결한다.

[0006] 제1 관점에 따르면, 성문 특징 모델 갱신 방법은: 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하는 단계; 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하는 단계; 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하는 단계; 및 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신하는 단계를 포함한다.

[0007] 제1 관점의 제1 가능한 실행 방법에서, 상기 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하는 단계 이전에, 상기 방법은: 사전설정된 오디오 스트림 학습 샘플에 따라 원래의 성문 특징 모델을 확립하는 단계를 더 포함한다.

[0008] 제1 관점 또는 제1 관점의 제1 가능한 실행 방법을 참조하여, 제2 가능한 실행 방법에서, 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하는 단계는 구체적으로: 상기 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하는 단계 - 상기 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함함 - ; 및 사전설정된 화자 클러스터링 알고리즘에 따라, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성하는 단계를 포함한다.

[0009] 제1 관점 또는 제1 관점의 제1 가능한 실행 방법 또는 제1 관점의 제2 가능한 실행 방법을 참조하여, 제3 가능한 실행 방법에서, 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하는 단계는: 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 상기 원래의 성문 특징 모델에 따라, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 획득하는 단계; 및 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 상기 성공적으로 정합된 오디오 스트림으로 선택하는 단계를 포함한다.

[0010] 제1 관점 또는 제1 관점의 제1 가능한 실행 방법 또는 제1 관점의 제2 가능한 실행 방법 또는 제1 관점의 제3 가능한 실행 방법을 참조하여, 제4 가능한 실행 방법에서, 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신하는 단계는 구체적으로: 상기 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하는 단계 - 상기 사전설정된 오디오 스트림 학습 샘플은 상기 원래의 성문 특징 모델을 생성하기 위한 오디오 스트림임 - ; 및 상기 정정된 성문 특징 모델에 따라 상기 원래의 성문 특징 모델을 갱신하는 단계를 포함한다.

[0011] 제2 관점에 따라 단말은, 원래의 오디오 스트림 획득 유닛, 세그먼테이션 및 클러스터링 유닛, 정합 유닛, 및 모델 갱신 유닛을 포함하며, 여기서 상기 원래의 오디오 스트림 획득 유닛은 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하고 상기 원래의 오디오 스트림을 상기 세그먼테이션 및 클러스터링 유닛에 송신하도록 구성되어 있으며; 상기 세그먼테이션 및 클러스터링 유닛은 상기 원래의 오디오 스트림 획득 유닛에 의해 송신된 원래의 오디오 스트림을 수신하고, 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하며, 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 상기 정합 유닛에 송신하도록 구성되어 있으며; 상기 정합 유닛은 상기 세그먼테이션 및 클러스터링 유닛에 의해 송신된 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 수신하고, 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하며, 상기 성공적으로 정합된 오디오 스트림을 상기 모델 갱신 유닛에 송신하도록 구성되어 있으며; 그리고 상기 모델 갱신 유닛은 상기 모델 갱신 유닛에 송신된 상기 성공적으로 정합된 오디오 스트림을 수신하고, 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하며, 상기 원래의 성문 특징 모델을 갱신하도록 구성되어 있다.

[0012] 제2 관점의 제1 가능한 실행 방식에서, 단말은 샘플 획득 유닛 및 원래의 모델 확립 유닛을 더 포함하며, 상기 샘플 획득 유닛은 사전설정된 오디오 스트림 학습 샘플을 획득하고, 상기 사전설정된 오디오 스트림 학습 샘플을 상기 원래의 모델 확립 유닛에 송신하도록 구성되어 있으며, 상기 원래의 모델 확립 유닛은 상기 샘플 획득 유닛에 의해 송신된 상기 사전설정된 오디오 스트림 학습 샘플을 수신하고, 상기 사전설정된 오디오 스트림 학습 샘플에 따라 상기 원래의 성문 특징 모델을 확립하도록 구성되어 있다.

[0013] 제2 관점 또는 제2 관점의 제1 가능한 실행 방식을 참조해서, 제2 가능한 실행 방식에서, 상기 세그먼테이션 및 클러스터링 유닛은 구체적으로 세그먼테이션 유닛 및 클러스터링 유닛을 포함하며, 여기서 상기 세그먼테이션 유닛은 상기 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하며 - 상기 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함함 - 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 상기 클러스터링 유닛에 송신하도록 구성되어 있으며; 그리고 상기 클러스터링 유닛은 상기 세그먼테이션 유닛에 의해 송신된, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 수신하고, 사전설정된 화자 클러스터링 알고리즘에 따라, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성하도록 구성되어 있다.

[0014] 제2 관점 또는 제2 관점의 제1 가능한 실행 방식 또는 제2 관점의 제2 가능한 실행 방식을 참조해서, 제3 가능한 실행 방식에서, 상기 정합 유닛은 구체적으로 정합도 획득 유닛 및 정합 오디오 스트림 획득 유닛을 포함하며, 여기서 상기 정합도 획득 유닛은 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 상기 원래의 성문 특징 모델에 따라, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 획득하고, 상기 정합도를 상기 정합 오디오 스트림 획득 유닛에 송신하도록 구성되어 있으며; 그리고 상기 정합 오디오 스트림 획득 유닛은 상기 정합도 획득 유닛에 의해 송신된, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 수신하고, 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 상기 성공적으로 정합된 오디오 스트림으로 선택하도록 구성되어 있다.

[0015] 제2 관점 또는 제2 관점의 제1 가능한 실행 방식 또는 제2 관점의 제2 가능한 실행 방식 또는 제2 관점의 제3 가능한 실행 방식을 참조해서, 제4 가능한 실행 방식에서, 상기 모델 갱신 유닛은 구체적으로 정정 모델 획득 유닛 및 모델 갱신 서브유닛을 포함하며, 여기서 상기 정정 모델 획득 유닛은 상기 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하고, 상기 정정된 성문 특징 모델을 상기 모델 갱신 서브유닛에 송신하도록 구성되어 있으며; 그리고 상기 모델 갱신 서브유닛은 상기 정정 모델 획득 유닛에 의해 송신된 상기 정정된 성문 특징 모델을 수신하고, 상기 정정된 성문 특징 모델에 따라 상기 원래의 성문 특징 모델을 갱신하도록 구성되어 있다.

## *발명의 효과*

[0016] 본 발명의 실시예에서는, 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하고, 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 사전설정된 화자 세그먼테이션

및 클러스터링 알고리즘에 따라 획득하며, 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림을 원래의 성문 특징 모델과 개별적으로 정합하여 성공적으로 정합된 오디오 스트림을 획득하며, 그 성공적으로 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로 사용함으로써, 원래의 성문 특징 모델을 갱신할 수 있다. 이것은 기존의 방법을 사용하여 성문 특징 모델을 획득할 때, 상대적으로 높은 실용성의 전제 하에서는 성문 특징 모델의 정확도를 개선하는 것이 확보될 수 없고, 그 결과 성문 특징 모델을 사용해서는 인식 정확도를 개선할 수 없다는 문제를 해결하며, 이것은 성문 특징 모델의 정확도 및 인식 정확도를 개선한다.

## *도면의 간단한 설명*

[0017] 도 1은 본 발명의 실시예 1에 따른 성문 특징 모델 갱신 방법을 실행하는 흐름도이다.

도 2는 본 발명의 실시예 2에 따른 성문 특징 모델 갱신 방법을 실행하는 흐름도이다.

도 3은 본 발명의 실시예에 따른 원래의 오디오 스트림의 세그먼테이션 및 클러스터링에 대한 개략도이다.

도 4는 본 발명의 실시예 3에 따른 단말에 대한 구조도이다.

도 5는 본 발명의 실시예 4에 따른 단말에 대한 구조도이다.

도 6은 본 발명의 실시예 5에 따른 단말에 대한 구조도이다.

도 7은 본 발명의 실시예 6에 따른 단말에 대한 구조도이다.

## *발명을 실시하기 위한 구체적인 내용*

[0018] 본 발명의 실시예의 목적, 기술적 솔루션, 및 이점을 더 명확하고 더 잘 이해할 수 있도록 하기 위해, 이하에서는 본 발명의 실시예의 첨부된 도면을 참조하여 본 발명의 실시예에 따른 기술적 솔루션에 대해 명확하고 완전하게 설명한다. 당연히, 이하의 상세한 설명에서의 실시예는 본 발명의 모든 실시예가 아닌 일부에 지나지 않는다. 당업자가 창조적 노력 없이 본 발명의 실시예에 기초하여 획득하는 모든 다른 실시예는 본 발명의 보호 범위 내에 있게 된다.

[0019] 본 발명의 실시예에서는, 적어도 한 명의 화자의 원래의 오디오 스트림을 획득하고, 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 획득하며, 원래의 성문 특징 모델과 정합하는 오디오 스트림을 획득하며, 그 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로 사용함으로써, 원래의 성문 특징 모델을 갱신할 수 있으며, 이에 따라 성문 특징 모델의 정확도를 개선하고 사용자 경험에 대한 효과가 향상된다.

[0020] 이하 본 발명의 특정한 실행에 대해 특정한 실시예를 참조하여 설명한다.

[0021] **실시예 1:**

[0022] 도 1은 본 발명의 실시예 1에 따른 성문 특징 모델 갱신 방법을 실행하는 프로세스이고 다음과 같이 상세히 설명한다:

[0023] 단계 S101: 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득한다.

[0024] 원래의 오디오 스트림은 사용자가 이동 단말을 사용하여 전화를 걸거나 음성 채팅을 함으로써 생성하는 오디오 스트림일 수 있으며, 또는 음성을 레코딩하는 방식으로 획득된 오디오 스트림일 수 있다. 구체적으로, 다음과 같은 상황이 가능하다: 특정한 이동 단말 사용자가 호출 접속 상태에 있을 때, 사용자가 성문 학습 기능을 사용하는 것에 동의하는지를 질의하고, 동의하면 대화 동안 생성된 오디오 스트림이 기록되거나; 성문 학습 기능을 자동으로 가능하게 하는 스위치가 호출 동안 단말에 대해 구성되고, 사용자가 필요에 따라 스위치를 설정하거나; 성문 학습 기능이 단말에 대해 구성되고, 사용자는 오디오 스트림을 기록할 수 있다. 여러 사람이 전화 동안 또는 채팅 동안 차례로 대화에 참여할 수 있으므로, 이 경우에 획득된 원래의 오디오 스트림은 여러 사람의 오디오 데이터를 포함할 수 있다는 것에 유의해야 한다.

[0025] 단계 S102: 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득한다.

[0026] 구체적으로, 원래의 오디오 스트림은 적어도 한 명의 화자의 오디오 스트림을 포함하고 있기 때문에, 이 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할해야 하며, 여기서 복수의 오디오 클립 중 각각의 오디오 클립은 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함한다. 그런 다음 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립은 사전설정된 화자 클러스터링 알고리즘에 따라 클러스터링되어 최종적으로 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성한다.

[0027] 단계 S103: 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득한다.

[0028] 원래의 성문 모델은 사전설정된 오디오 스트림 학습 샘플에 따라 미리 확립되어 있는 성문 특징 모델이다. 원래의 성문 특징 모델은 특정한 사람이나 여러 사람에 대한 성문 등록 프로세스 후에 형성되는 특징 모델이고, 이 등록 프로세스는 학습 말뭉치의 길이에 대한 요건을 가지지 않으며, 이를 오디오 스트림 학습 샘플이라고도 한다. 이 경우, 성공적으로 정합된 오디오 스트림은 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도(matching degree)에 따라 선택될 수 있다.

[0029] 단계 S104: 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신한다.

[0030] 구체적으로, 상기 성공적으로 정합된 오디오 스트림을 획득한 후, 상기 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플을 기본으로 사용하며, 여기서 사전설정된 오디오 스트림 학습 샘플은 전술한 원래의 성문 특징 모델을 생성하기 위한 샘플이다. 그런 다음, 성문 등록 알고리즘 인터페이스를 호출하고, 정정된 성문 특징 모델이 생성되며, 여기서 정정된 성문 특징 모델은 더 정확한 성문 특징 모델이며, 이에 의해 모델 적응 및 지능의 목적을 달성한다.

[0031] 선택적으로, 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림이 원래의 성문 특징 모델과 정합할 수 없는 상황에서, 성문 특징 모델은 사용자의 사전-설정에 따라 새롭게 확립되고 기록될 수 있다. 예를 들어, 처음 사용되는 단말에 있어서는, 원래의 성문 특징 모델이 무효이고, 정합에 사용되는 오디오 스트림은 존재하지 않는다. 이 경우, 특정한 스피커의 오디오 스트림은 사용자의 설정에 따라 인식되고, 성문 등록 알고리즘 인터페이스는 성문 특징 모델을 새롭게 확립하기 위해 호출되고, 원래의 성문 특징 모델은 그 새롭게 확립된 성문 특징 모델로 갱신된다.

[0032] 본 발명의 본 실시예에서는, 적어도 한 명의 화자의 원래의 오디오 스트림을 획득하고, 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 획득하며, 원래의 성문 특징 모델과 정합하는 오디오 스트림을 획득하며, 그 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하며, 원래의 성문 특징 모델을 갱신하며, 이에 의해 성문 특징 모델을 지속적으로 정정하고 갱신하는 목적을 달성하고, 성문 특징 모델의 정확도를 지속적으로 높이며, 사용자 경험을 향상시키는 등을 달성할 수 있다.

[0033] **실시예 2:**

[0034] 도 2는 본 발명의 실시예 2에 따른 성문 특징 모델 갱신 방법을 실행하는 프로세스이고 다음과 같이 상세히 설명한다:

[0035] 단계 S201: 사전설정된 오디오 스트림 학습 샘플에 따라 원래의 성문 특징 모델을 확립한다.

[0036] 원래의 성문 특징 모델은 성문 등록 알고리즘 인터페이스를 호출함으로써 사전설정된 오디오 스트림 학습 샘플에 따라 확립된 성문 특징 모델이다.

[0037] 원래의 성문 특징 모델은 특정한 사람이나 여러 사람에 대한 성문 등록 프로세스 후에 형성되는 특징 모델이고, 이 등록 프로세스는 학습 말뭉치의 길이에 대한 요건을 가지지 않으며, 이를 오디오 스트림 학습 샘플이라고도 한다. 또한, 본 발명의 실시예에서 제공하는 방법은 정정된 모델에 대해 동적 정정을 지속적으로 수행할 수 있고, 원래의 성문 특징 모델은 기존의 방법을 사용하여 획득된 모델일 수 있으며, 본 발명의 본 실시예에서 제공하는 방법을 사용하여 정정된 모델일 수도 있다.

[0038] 단계 S202: 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득한다.

[0039] 특정한 실행 프로세스에서, 원래의 오디오 스트림은

[0040] 사용자가 이동 단말을 사용하여 전화를 걸거나 음성 채팅을 함으로써 생성하는 오디오 스트림일 수 있으며, 또는 음성을 레코딩하는 방식으로 획득된 오디오 스트림일 수 있다. 구체적으로, 다음과 같은 상황이 가능하다: 특정한 이동 단말 사용자가 호출 접속 상태에 있을 때, 사용자가 성문 학습 기능을 사용하는 것에 동의하는지를 질의하고, 사용자가 동의한 후, 대화 동안 생성된 오디오 스트림이 기록되거나; 호출 동안 성문 학습 기능을 자동으로 가능하게 하는 스위치가 단말에 대해 구성되고, 사용자는 필요에 따라 스위치를 설정하거나; 성문 학습 기능이 단말에 대해 구성되고, 사용자는 오디오 스트림을 기록할 수 있다. 통상적으로 여러 사람이 전화 동안 또는 채팅 동안 차례로 대화에 참여할 수 있으므로, 이 경우에 획득된 원래의 오디오 스트림은 여러 사람의 오디오 데이터를 포함할 수 있다는 것에 유의해야 한다.

[0041] 또한, 말하는 속도, 억양, 및 감정 동요는 사용자가 말하는 프로세스 또는 여러 사람의 대화의 프로세스 동안 크게 변할 수 있다. 호출 동안의 말뭉치는 성문 특징 모델의 정확도를 위해 사람의 억양, 말하는 속도, 및 감정의 요인으로 야기되는 편차를 제거하도록 지속적으로 수집되고, 이것은 성문 특징 모델의 정확도에 대한 억양, 말하는 속도, 및 감정의 요인의 영향을 크게 감소시키고, 또한 성문 인식 정확도에 대한 충격을 감소시킬 수 있다.

[0042] 단계 S203: 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하고, 여기서 복수의 오디오 스트림의 각각의 오디오 클립은 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함한다.

[0043] 단계 S204: 사전설정된 화자 클러스터링 알고리즘에 따라, 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성한다.

[0044] 구체적으로, 여러 사람의 대화를 예로 해서, 대화에 참여하고 있는 사람이 사용자 A, 사용자 B, 사용자 C인 것으로 가정한다. 사용자가 음성을 레코딩하는 것에 동의한 후, 레코딩 모듈이 가능하게 되고, 호출 동안의 원래의 오디오 스트림은 호출이 완료된 후 또는 레코딩 지속시간이 만료된 후 레코딩된다. 원래의 오디오 스트림은 원래의 오디오 스트림은 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할될 수 있으며, 여기서 각각의 오디오 클립은 한 명의 화자의 오디오 정보만을 포함한다. 도 3에 도시된 바와 같이, 원래의 오디오 스트림이 분할된 후, 그 획득된 오디오 클립은 오디오 클립 A, 오디오 클립 B, 오디오 클립 A, 오디오 클립 C, 오디오 클립 A, 오디오 클립 C이며; 오디오 클립 A, 오디오 클립 B, 오디오 클립 C는 각각 사용자 A, B, C로 되어 있는 상이한 클립이며, 말하는 순번에 따라 획득된다. 그런 다음, 동일한 화자의 오디오 클립은 사전설정된 화자 클러스터링 알고리즘을 사용하여 클러스터링됨으로써, 오디오 스트림 A의 파일, 오디오 스트림 B의 파일, 및 오디오 스트림 C의 파일을 생성한다. 예를 들어, 오디오 스트림 A는 사용자 A의 모든 오디오 클립이다. 그러므로 상이한 사람들의 오디오 스트림을 구별할 수 있고, 동일한 사람의 유효한 오디오 스트림을 추출할 수 있다. 화자 세그먼테이션 알고리즘 및 클러스터링 알고리즘은 각각 기존의 임의의 한 명의 화자 세그먼테이션 알고리즘 및 클러스터링 알고리즘일 수 있으며, 여기서 제한되지 않는다.

[0045] 단계 S205: 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득한다.

[0046] 단계 S205는 구체적으로:

[0047] 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 상기 원래의 성문 특징 모델에 따라, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도(matching degree)를 획득하는 단계; 및

[0048] 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 상기 성공적으로 정합된 오디오 스트림으로 선택하는 단계

[0049] 를 포함한다.

[0050] 구체적으로, 성문 증명 알고리즘 인터페이스를 호출하여, 오디오 스트림 A, 오디오 스트림 B, 오디오 스트림 C와 원래의 성문 특징 모델 간의 정합도 A, 정합도 B, 정합도 C를 개별적으로 획득한다. 정합도의 계산 방식은: 오디오 스트림 A, 오디오 스트림 B, 오디오 스트림 C를 각각 원래의 성문 특징 모델의 입력값으로 사용하고, 원래의 성문 특징 모델에 대응하여, 오디오 스트림 A, 오디오 스트림 B, 오디오 스트림 C의 정합도 A, 정합도 B, 정합도 C를 각각 획득하며, 여기서 정합도 A, 정합도 B, 정합도 C를 각각 대응하는 확률 A, 확률 B, 확률 C라고

도 한다. 예를 들어, 정합도 A는 오디오 스트림 A와 원래의 특징 모델 간의 관련성을 나타낸다. 원래의 성문 특징 모델은 사용자 A의 오디오 스트림 학습 샘플에 기초해서 구축되고, 정합도 A는 정상적인 상황 하에서 정합 임계값보다 크고, 정합도 B 및 정합도 C는 정상적인 상황 하에서는 정합 임계값보다 작아야 하는 것으로 가정하고, 여기서 사전설정된 임계값은 실제의 테스트 결과에 따라 획득될 수 있거나, 사전설정될 수 있거나, 사용자-정의일 수 있다. 그러므로 이 경우 사전설정된 임계값에 대응하는 정합도에 대응하는 오디오 스트림이 획득되며, 즉 오디오 스트림 A가 성공적으로 정합된 오디오 스트림이다. 특별한 경우, A 및 B의 사운드가 유사할 때, 정합 임계값보다 큰 하나 이상의 오디오 스트림이 있을 수 있으며, 정합 값이 가장 높은 오디오 스트림을 성공적으로 정합된 오디오 스트림으로 선택할 수 있다.

[0051] 또한, 원래의 성문 특징 모델이 여러 사람에 대한 성문 등록 프로세스 후에 형성되는 특징 모델일 때, 예를 들어 사용자 B 및 C의 오디오 스트림 학습 샘플에 대해 구축되는 특징 모델일 때, 그 정합 후에 획득되는 오디오 스트림은 모두 오디오 스트림 B 및 오디오 스트림 C를 포함할 가능성이 크고, 이에 의해 다인 모드(multi-person mode)에서 성문 특징 모델의 정합을 실행한다. 이 경우, 전술한 단계들은 여러 사람 각각에 대해 개별적으로 실행된다.

[0052] 단계 S206: 상기 성공적으로 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로 사용하고, 상기 원래의 성문 모델 특징을 갱신한다.

[0053] 단계 S206은 구체적으로:

[0054] 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하는 단계 - 상기 사전설정된 오디오 스트림 학습 샘플은 원래의 성문 특징 모델을 생성하기 위한 오디오 스트림임 - ; 및

[0055] 상기 원래의 성문 특징 모델을 상기 정정된 성문 특징 모델로 갱신하는 단계

[0056] 를 포함한다.

[0057] 구체적으로, 성공적으로 정합된 오디오 스트림은 추가의 오디오 스트림 학습 샘플로서 사용된다. 즉, 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하기 위해 성문 등록 알고리즘 인터페이스를 호출하며, 여기서 상기 정정된 성문 특징 모델은 더 정확한 성문 특징 모델이며, 이에 의해 모델 적응 및 지능의 목적을 달성한다.

[0058] 또한, 갱신된 성문 특징 모델은 원래의 성문 모델로도 사용될 수 있으며, 전술한 단계들은 지속적으로 반복되어 성문 특징 모델을 정정하고 갱신하며, 성문 특징 모델의 정확도를 지속적으로 높인다.

[0059] 본 발명의 본 실시예에서, 음성 호출의 원래의 오디오 스트림은 자동으로 성문 학습 말뭉치로 사용되고, 수집된 원래의 오디오 스트림은 사용자 경험이 영향을 받지 않거나 사용자 조작이 감소되는 상황에서 화자 세그먼테이션 및 클러스터링 알고리즘을 사용해서 처리되어 성문 학습 말뭉치의 순수성을 보장하며, 추가의 정합된 오디오 스트림은 학습 말뭉치의 길이를 늘이는 데 사용되어, 원래의 성문 특징 모델을 동적으로 정정한다. 이것은 성문 특징 모델을 동적으로 정정하고 갱신하며 성문 특징 모델의 정확도를 높인다. 그러므로 인식률을 더 높일 수 있고 사용자의 사용자 경험 역시 성문 특징 모델을 사용함으로써 음성 인식과 같은 프로세스에서 개선될 수 있다.

[0060] 당업자라면 전술한 실시예의 방법의 단계 중 일부 또는 전부는 관련 하드웨어에 명령을 내리는 프로그램으로 실행될 수 있다는 것을 이해할 수 있을 것이다. 프로그램은 컴퓨터 판독 가능형 저장 매체에 저장될 수 있으며, 이러한 저장 매체로는 예를 들어 ROM/RAM, 자기디스크, 또는 광디스크를 들 수 있다.

[0061] **실시예 3:**

[0062] 도 4는 본 발명의 실시예 3에 따른 단말에 대한 구조도이다. 본 발명의 실시예 3에서 제공하는 단말은 본 발명의 실시예 1 및 실시예 2의 방법을 실행하도록 구성될 수 있다. 설명을 쉽게 하기 위해, 본 발명의 실시예와 관련된 부분만을 도시하고 있다. 설명되지 않은 특정한 기술적 상세에 대해서는, 본 발명의 실시예 1 및 실시예 2를 참조하면 된다.

[0063] 단말은 이동 전화, 태블릿 컴퓨터, 개인휴대단말(Personal Digital Terminal: PDA), 판매 시점 관리(Point of Sales: POS) 시스템, 또는 차량 장착 컴퓨터와 같은 단말 장치일 수 있다. 단말이 이동 전화인 경우를 예를 들어 사용한다. 도 4는 본 발명의 본 실시예에서 제공하는 단말과 관련된 이동 전화(400)의 구조 중 일부에 대한

블록도이다. 도 4를 참조하면, 이동 전화(400)는 무선 주파(Radio Frequency: RF) 회로(410), 메모리(420), 입력 유닛(430), 디스플레이 유닛(440), 센서(450), 오디오 회로(460), WiFi(wireless fidelity) 모듈(470), 프로세서(480), 전원(490)과 같은 부분을 포함한다. 당업자라면 도 4에 도시된 이동 전화의 구조는 이동 전화에 대한 제한을 구성하지 않으며, 이동 전화는 도면에 도시된 것보다 더 많은 또는 더 적은 부분을 포함할 수 있거나 일부의 부분을 결합할 수 있거나, 이러한 부분들의 상이한 배치를 가질 수 있다는 것을 이해할 수 있을 것이다.

[0064] 이하에서는 이러한 이동 전화(400)의 부분을 도 4를 참조하여 상세히 설명한다.

[0065] RF 회로(410)는 정보를 송수신하거나 호출 동안 신호를 송수신하며, 특히 기지국의 다운링크 정보를 수신하고, 그 정보를 처리를 위한 프로세서(480)에 송신한다. 또한, RF 회로(410)는 업링크 데이터를 기지국에 송신한다. 일반적으로, RF 회로는 적어도 하나의 증폭기, 송수신기, 커플러, 저잡음 증폭기(Low Noise Amplifier: LNA), 듀플렉서 등을 포함하되, 이에 제한되지 않는다. 또한, RF 회로(410)는 무선 통신 및 네트워크를 사용하여 다른 장치들과 통신할 수 있다. 무선 통신은 임의의 하나의 통신 표준 또는 프로토콜을 사용할 수 있으며, 이러한 통신 표준 또는 프로토콜로는 이동 통신을 위한 글로벌 시스템(Global System for Mobile Communication: GSM), 범용 패킷 무선 서비스(General Packet Radio Service: GPRS), 코드분할다중접속(Code Division Multiple Access: CDMA), 광대역 코드분할다중접속(Wideband Code Division Multiple Access: WCDMA), 롱텀에볼루션(Long Term Evolution: LTE), 전자 메일, 단문 메시징 서비스(Short Messaging Service: SMS) 등이 있으나, 이에 제한되지 않는다.

[0066] 메모리(420)는 소프트웨어 프로그램 및 모듈을 저장하도록 구성될 수 있다. 프로세서(480)는 메모리(420)에 저장되어 있는 소프트웨어 프로그램 및 모듈을 실행하여 이동 전화(400)의 모든 유형의 기능 애플리케이션을 실행하고 데이터를 처리한다. 메모리(420)는 프로그램 저장 영역 및 데이터 저장 영역을 주로 포함할 수 있으며, 여기서 프로그램 저장 영역은 운영체제, 기능이 필요로 하는 적어도 하나의 애플리케이션 프로그램(예를 들어, 사운드 재생 기능 또는 이미지 재생 기능 등) 등을 저장할 수 있으며; 데이터 저장 영역은 이동 전화(400)의 사용에 따라 생성되는 데이터(예를 들어, 오디오 데이터 또는 전화번호부)를 저장할 수 있다. 또한, 메모리(420)는 고속의 랜덤 액세스 메모리를 포함할 수 있고, 비휘발성 메모리, 예를 들어, 적어도 하나의 자기디스크, 플래시 메모리 또는 다른 비휘발성 고체상태 메모리도 포함할 수 있다.

[0067] 입력 유닛(430)은 입력되는 숫자 또는 문자 정보를 수신하고 이동 전화(400)의 사용자 설정 및 기능 제어와 관련된 중요한 신호 입력을 생성하도록 구성될 수 있다. 구체적으로, 입력 유닛(430)은 터치-제어 패널(431) 및 다른 입력 장치(432)를 포함할 수 있다. 터치-제어 패널(431)은 터치 스크린이라고도 하는데 패널 상에서 또는 패널 근처에서의 사용자의 터치 동작(예를 들어, 사용자가 손가락 또는 스트일러스와 같은 임의의 적절한 대상 또는 부착물을 사용하여 터치-제어 패널(431) 상에서 또는 터치-제어 패널(431) 근처에서 수행하는 동작)을 수집할 수 있고, 미리 정해진 프로그램에 따라 대응하는 접속 장치를 구동시킬 수 있다. 선택적으로, 터치-제어 패널(431)은 2개의 부분, 즉 터치 검출 장치 및 터치 제어기를 포함할 수 있다. 터치 검출 장치는 사용자의 터치 위치를 검출하고, 터치 동작에 의해 생기는 신호를 검출하며, 이 신호를 터치 제어기에 전달한다. 터치 제어기는 터치 검출 장치로부터 터치 정보를 수신하고, 이 터치 정보를 터치 포인트의 좌표로 변환하고, 이 좌표를 프로세서(480)에 송신하며, 프로세서(480)에 의해 송신되는 커맨드를 수신 및 실행할 수 있다. 또한, 터치-제어 패널(431)은 복수의 형태, 예를 들어, 저항성 형태, 용량성 형태, 및 표면 음향 파로 실행될 수 있다. 터치-제어 패널(431) 외에도, 입력 유닛(430)은 다른 입력 장치(432)를 더 포함할 수 있다. 구체적으로, 다른 입력 장치(432)는 물리적 키보드, 기능 키(예를 들어 음량 제어 키 및 온-오프 키), 트랙볼, 마우스, 및 조이스틱 중 하나 이상을 포함하되, 이에 제한되지는 않는다.

[0068] 디스플레이 유닛(440)은 사용자가 입력하는 정보, 사용자에게 제공되는 정보, 및 이동 전화(400)의 다양한 메뉴를 표시하도록 구성될 수 있다. 디스플레이 유닛(440)은 디스플레이 패널(441)을 포함할 수 있다. 선택적으로, 디스플레이 패널(441)은 액정 디스플레이(Liquid Crystal Display: LCD) 및 유기발광 다이오드(Organic Light-Emitting Diode: OLED) 등의 형태를 사용하여 구성될 수 있다. 또한, 터치-제어 패널(431)은 디스플레이 패널(441)을 커버할 수 있다. 터치-제어 패널(431) 상의 또는 터치-제어 패널(431) 근처의 터치 동작을 검출한 후, 터치-제어 패널(431)은 이 동작을 프로세서(480)에 전송하여 터치 이벤트의 유형을 판단한다. 그런 다음 프로세서(480)는 그 터치 이벤트의 유형에 따라 대응하는 비주얼 출력을 디스플레이 패널(441) 상에 제공한다. 도 4의 터치-제어 패널(431) 및 디스플레이 패널(441)이 이동 전화(400)의 입력 및 입력 기능을 실행하는 2개의 독립적인 구성요소이나, 일부의 실시예에서는 터치-제어 패널(431) 및 디스플레이 패널(441)을 통

합하여 이동 전화(400)의 입력 및 입력 기능을 실행할 수 있다.

[0069] 이동 전화(400)는 적어도 하나의 유형의 센서(450), 예를 들어, 광센서, 모션 센서, 및 다른 센서를 더 포함할 수 있다. 구체적으로, 광센서는 조도 센서(ambient light sensor) 및 근접 센서(proximity sensor)를 포함하고, 조도 센서는 조도의 밝기에 따라 디스플레이 패널(441)의 밝기를 표시하며, 근접 센서는 이동 전화(400)를 귀에 가까이 대면 디스플레이 패널(441) 및/또는 백라이트를 턴 오프할 수 있다. 한 유형의 모션 센서로서, 가속 센서는 모든 방향(일반적으로 3축)의 가속을 검출할 수 있고, 가속도계가 정적 상태에 있을 때 중력의 크기 및 방향을 검출할 수 있으며, 이동 단말 자세 애플리케이션(예를 들어, 초상화와 풍경 방향 간의 전환, 관련 게임, 및 자력계 자세 캘리브레이션), 및 진동 인식 관련 기능(예를 들어, 계수기 및 두드림)을 인식하도록 구성될 수 있다. 자이로스코프, 기압계, 습도계, 온도계, 및 적외선 센서와 같은 다른 센서가 이동 전화(400)에 구성될 수 있으며, 이에 대해서는 여기서 더 설명하지 않는다.

[0070] 오디오 회로(460), 라우드스피커(461), 및 마이크로폰(462)은 사용자와 이동 전화(400) 간에 오디오 인터페이스를 제공할 수 있다. 오디오 회로(460)는 수신된 오디오 데이터로부터 변환된 전기 신호를 변환하여 라우드스피커(461)에 전송하고, 라우드스피커(461)는 이 전기 신호를 음성 신호로 변환하여 출력한다. 한편, 마이크로폰(462)은 수집된 사운드 신호를 전기 신호로 변환하고, 오디오 회로(460)는 전기 신호를 수신하고 이 전기 신호를 오디오 데이터로 변환하고, 이 오디오 데이터를 처리를 위한 프로세서(480)에 출력하며, 처리된 오디오 신호를, 예를 들어, RF 회로(410)를 사용함으로써 다른 이동 전화에 송신되거나, 이 오디오 데이터를 추가의 처리를 위해 메모리(420)에 출력한다.

[0071] WiFi는 단거리 무선 전송 기술에 속한다. 이동 전화(400)는 WiFi 모듈(470)을 사용함으로써, 사용자가 전자메일을 송수신할 수 있게 하고, 웹페이지를 브라우징할 수 있게 하며, 스트리밍 미디어에 액세스할 수 있게 한다. WiFi 모듈(470)은 무선 광대역 인터넷 액세스를 사용자에게 제공한다. 도 4에는 WiFi 모듈(470)이 도시되어 있으나, WiFi 모듈은 이동 전화(400)의 필수 구성요소가 아니며 당연히 본 발명의 본질을 변화시킴이 없는 범위 내의 요구에 따라 생략될 수도 있다는 것을 이해할 수 있어야 한다.

[0072] 프로세서(480)는 이동 전화(400)의 제어 센터이며, 모든 유형의 인터페이스 및 회로를 사용함으로써 전체 이동 전화의 모든 부분을 접속하며, 메모리(420)에 저장되어 있는 소프트웨어 프로그램 및/또는 모듈을 운영 또는 실행함으로써 그리고 메모리(420)에 저장되어 있는 데이터를 호출함으로써 이동 전화를 전반적으로 모니터링한다. 선택적으로, 프로세서(480)는 하나 이상의 프로세싱 유닛을 포함할 수 있다. 양호하게, 프로세서(480)는 애플리케이션 프로세서 및 모뎀 프로세서와 통합될 수 있으며, 여기서 애플리케이션 프로세서는 운영체제, 사용자 인터페이스, 애플리케이션 프로그램 등을 주로 처리하며, 모뎀 프로세서는 무선 통신을 주로 처리한다. 전술한 모뎀 프로세서는 프로세서(480)에 통합되지 않을 수도 있음은 물론이다.

[0073] 이동 전화(400)는 모든 구성요소에 전력을 공급하는 전원(490)(예를 들어, 배터리)을 더 포함한다. 양호하게, 전원은 전원 관리 시스템을 사용함으로써 프로세서(480)에 논리적으로 접속되어 있고, 그러므로 전원 관리 시스템을 사용함으로써 충전 관리, 방전 관리, 및 전력 소모 관리와 같은 기능을 실행한다.

[0074] 도시되지는 않았으나, 이동 전화(400)는 카메라 및 블루투스 모듈 등을 더 포함할 수 있으며, 이에 대해서는 여기서 더 설명하지 않는다.

[0075] 본 발명의 본 실시예에서, 단말에 의해 포함되는 마이크로폰(462), 메모리(420), 및 프로세서(480)는 이하의 기능을 더 구비한다.

[0076] 마이크로폰(462)은 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하고, 오디오 회로(460)를 사용하여 원래의 오디오 스트림을 메모리(420)에 송신하도록 구성되어 있다.

[0077] 본 발명의 본 실시예에서, 원래의 오디오 스트림은 사용자가 이동 단말을 사용하여 전화를 걸거나 음성 채팅을 함으로써 생성하는 오디오 스트림일 수 있으며, 또는 예를 들어 음성을 레코딩하는 방식으로 마이크로폰(462)에 의해 획득된 오디오 스트림일 수 있다. 구체적으로, 다음과 같은 상황이 가능하다: 특정한 이동 전화 단말이 호출 접속 상태에 있을 때, 사용자가 성문 학습 기능을 사용하는 것에 동의하는지를 질의하고, 사용자가 동의하면 대화 동안 생성된 오디오 스트림이 기록되거나; 호출 동안 성문 학습 기능을 자동으로 가능하게 하는 스위치가 단말에 대해 구성되고, 사용자가 필요에 따라 스위치를 설정하거나; 성문 학습 기능이 이동 전화 단말에 대해 구성되고, 사용자는 오디오 스트림을 기록할 수 있다. 여러 사람이 전화 동안 또는 채팅 동안 차례로 대화에 참여할 수 있으므로, 이 경우에 획득된 원래의 오디오 스트림은 여러 사람의 오디오 데이터를 포함할 수 있다는 것에 유의해야 한다.

[0078] 프로세서(480)는 메모리에 저장되어 있는 원래의 오디오 스트림을 발동하고, 메모리(420) 내의 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘을 호출하고, 이 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하며, 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 원래의 성문 특징 모델과 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하고, 이 성공적으로 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 원래의 성문 특징 모델을 갱신한다.

[0079] 본 발명의 본 실시예에서, 원래의 오디오 스트림은 적어도 한 명의 화자의 오디오 스트림을 포함하고 있기 때문에, 프로세서(480)는 메모리(420) 내의 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘을 호출하고 원래의 오디오 스트림을 복수의 오디오 클립으로 분할하며 여기서, 상기 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함한다. 그런 다음, 프로세서(480)는 사전설정된 화자 클러스터링 알고리즘에 따라, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 최종적으로 생성한다. 또한, 프로세서(480)는 각각의 사람의 각각의 오디오 스트림 및 원래의 성문 특징 모델을 참조하여 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합함으로써 획득되는 정합도를 획득할 수 있고,

[0080] 사전설정된 정합 임계값보다 크면서 가장 높은 정합도를 가지는 오디오 스트림을 성공적으로 정합된 오디오 스트림으로 선택함으로써, 성공적으로 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고; 성문 등록 알고리즘 인터페이스를 호출하고 원래의 성문 특징 모델을 갱신하여, 더 정확한 성문 특징 모델을 획득한다.

[0081] 본 발명의 본 실시예는 마이크로폰(462), 메모리(420), 프로세서(480) 등을 포함하는 단말을 제공한다. 마이크로폰(462)은 적어도 한 명의 화자의 원래의 오디오 스트림을 획득하고, 이 원래의 오디오 스트림을 오디오 회로(460)를 통해 메모리(420)에 송신한다. 프로세서(480)는 오디오 회로(460)를 통해 마이크로폰(462)에 의해 송신되는 원래의 오디오 스트림을 수신하고, 메모리(420) 내의 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘을 호출하고, 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하며, 원래의 성문 특징 모델과 정합하는 오디오 스트림을 획득하며, 이 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하며, 원래의 성문 특징 모델을 갱신한다. 이것은 상대적으로 높은 실용성의 전제 하에 성문 특징 모델에 대한 동적 정정 및 갱신을 보장하고 성문 특징 모델의 정확도를 높인다.

[0082] **실시예 4:**

[0083] 도 5는 본 발명의 실시예 4에 따른 단말에 대한 구조도이다. 설명을 쉽게 하기 위해, 본 발명의 본 실시예와 관련된 부분만을 도시하고 있다. 본 발명의 실시예 4에서 제공하는 단말은 본 발명의 실시예 1 및 실시예 2의 방법을 실행하도록 구성될 수 있다. 설명을 쉽게 하기 위해, 본 발명의 본 실시예와 관련된 부분만을 도시하고 있다. 설명되지 않은 특정한 기술적 상세에 대해서는, 본 발명의 실시예 1 및 실시예 2를 참조하면 된다.

[0084] 구체적으로, 도 5는 본 발명의 본 실시예에서 제공하는 단말과 관련된 이동 전화(500)의 구조 중 일부에 대한 블록도를 도시한다. 도 4에 도시된 구조에 기초해서, 본 발명의 실시예에서는, 도 4에 도시된 마이크로폰(462) 및 프로세서(480) 대신 마이크로폰(51) 및 프로세서(52)를 각각 사용한다.

[0085] 실시예 3에서의 마이크로폰(462)에 포함되어 있는 기능 외에, 마이크로폰(51)은 사전설정된 오디오 스트림 학습 샘플을 획득하고, 오디오 회로(460)를 사용함으로써 오디오 스트림 학습 샘플을 메모리(420)에 송신하도록 구성되어 있으며, 이에 따라 프로세서(52)는 메모리 내의 사전설정된 성문 등록 알고리즘 인터페이스를 호출하고, 사전설정된 오디오 스트림 학습 샘플에 따라 원래의 성문 특징 모델을 확립한다.

[0086] 본 발명의 본 실시예에서, 원래의 성문 특징 모델은 성문 등록 알고리즘 인터페이스를 호출함으로써 사전설정된 오디오 스트림 학습 샘플에 따라 확립된 성문 특징 모델이다. 원래의 성문 특징 모델은 특정한 사람이나 여러 사람에 대한 성문 등록 프로세스 후에 형성되는 특징 모델이고, 이 등록 프로세스는 학습 말뭉치의 길이에 대한 요건을 가지지 않으며, 이를 오디오 스트림 학습 샘플이라고도 한다. 또한, 본 발명의 실시예에서 제공하는 방법은 정정된 모델에 대한 지속적이고 동적인 정정을 실행할 수 있고, 원래의 성문 특징은 기존의 방법을 사용함으로써 획득되는 모델일 수도 있고 본 발명의 실시예에서 제공하는 방법을 사용함으로써 정정된 모델일 수도 있다.

[0087] 이 경우, 프로세서(52)는, 적어도 한 명의 화자가 말을 할 때 마이크로폰(51)에 의해 수신되는 원래의 오디오 스트림에 따라, 메모리(420) 내의 사전설정된 스피커 세그먼테이션 알고리즘을 호출함으로써 원래의 오디오 스트림을 복수의 오디오 클립을 분할하고, 여기서 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하며, 그런 다음 메모리(420) 내의 사전설정된 화자 클러스터링 알고리즘을 호출함으로써, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성한다.

[0088] 또한, 프로세서(52)는 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 원래의 성문 특징 모델에 따라 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 획득하고, 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 성공적으로 정합된 오디오 스트림으로 선택하며, 상기 성공적으로 정합된 오디오 스트림 및 상기 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하며, 원래의 성문 특징 모델을 상기 정정된 성문 특징 모델로 갱신하도록 추가로 구성되어 있다.

[0089] 본 발명의 본 실시예에서, 마이크로폰(51)은 사전설정된 오디오 스트림 학습 샘플을 획득할 수 있으며, 여기서 사전설정된 오디오 스트림 학습 샘플은 원래의 성문 특징 모델을 확립하는 데 필요한 원래의 오디오 스트림이다. 마이크로폰(51)은 또한 적어도 한 명의 화자의 원래의 오디오 스트림을 획득할 수 있다. 프로세서(52)는 메모리(420) 내의 사전설정된 성문 등록 알고리즘 인터페이스, 화자 세그먼테이션 알고리즘, 및 사전설정된 화자 클러스터링 알고리즘을 선택적으로 호출하여, 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성하며, 최종적으로 성공적으로 정합된 오디오 스트림을 획득할 수 있으며; 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하고, 원래의 성문 특징 모델을 정정된 성문 특징 모델로 갱신한다. 그러므로 정정된 성문 특징 모델을 사용하여 원래의 성문 특징 모델에 비해 오디오 스트림 인식 정확도를 현저하게 높일 수 있고, 사용자 경험은 더 향상된다.

[0090] **실시예 5:**

[0091] 도 6은 본 발명의 실시예 5에 따른 단말에 대한 구조도이다. 설명을 쉽게 하기 위해, 본 발명의 본 실시예와 관련된 부분만을 도시하고 있다. 본 발명의 실시예 5에서 제공하는 단말은 본 발명의 실시예 1 및 실시예 2의 방법을 실행하도록 구성될 수 있다. 설명을 쉽게 하기 위해, 본 발명의 본 실시예와 관련된 부분만을 도시하고 있다. 설명되지 않은 특정한 기술적 상세에 대해서는, 본 발명의 실시예 1 및 실시예 2를 참조하면 된다.

[0092] 단말은 원래의 오디오 스트림 획득 유닛(61), 세그먼테이션 및 클러스터링 유닛(62), 정합 유닛(63), 및 모델 갱신 유닛(64)을 포함한다. 원래의 오디오 스트림 획득 유닛(61)은 실시예 3에서의 마이크로폰(41)에 의해 포함되어 있는 기능과 일대일 대응하고, 세그먼테이션 및 클러스터링 유닛(62), 정합 유닛(63) 및 모델 갱신 유닛(64)은 실시예 3에서의 프로세서(42)에 의해 포함되어 있는 기능들과 일대일 대응하며, 여기서,

[0093] 상기 원래의 오디오 스트림 획득 유닛(61)은 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득하고, 상기 원래의 오디오 스트림을 상기 세그먼테이션 및 클러스터링 유닛(62)에 송신하도록 구성되어 있으며;

[0094] 상기 세그먼테이션 및 클러스터링 유닛(62)은 상기 원래의 오디오 스트림 획득 유닛(61)에 의해 송신된 원래의 오디오 스트림을 수신하고, 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 상기 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득하며, 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 상기 정합 유닛(63)에 송신하도록 구성되어 있으며;

[0095] 상기 정합 유닛(63)은 상기 세그먼테이션 및 클러스터링 유닛(62)에 의해 송신된 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 수신하고, 상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득하며, 상기 성공적으로 정합된 오디오 스트림을 상기 모델 갱신 유닛(64)에 송신하도록 구성되어 있으며; 그리고

[0096] 상기 모델 갱신 유닛(64)은 상기 모델 갱신 유닛(63)에 송신된 상기 성공적으로 정합된 오디오 스트림을 수신하고, 상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하며, 상기 원래의 성문 특징 모델을 갱신하도록 구성되어 있다.

[0097] 본 발명의 본 실시예에서, 호출 듣기 상태(call listening state)로 들어간 후, 원래의 오디오 스트림 획득 유

닛(61)은 듣기에 의해 오디오 스트림을 획득하며 여기서 오디오 스트림은 음성 레코더 또는 음성 채팅 소프트웨어를 사용하여 생성될 수 있다.

[0098] 본 발명의 본 실시예에서, 상기 세그먼테이션 및 클러스터링 유닛(62)은 원래의 오디오 스트림을 수 개의 오디오 클립으로 분할 수 있고, 여기서 각각의 오디오 클립은 한 명의 화자의 오디오 정보만을 포함하며, 동일한 화자의 오디오 클립을 다시 클러스터링하여 각자의 오디오 스트림을 생성하며, 최종적으로 원래의 오디오 스트림을 상이한 화자를 나타내는 오디오 스트림으로 분할하고, 즉 모든 화자 중에서 동일한 화자의 오디오 정보의 오디오 스트림을 생성한다. 정합 회로(63)는 모든 오디오 스트림을 횡단하고, 원래의 성문 특징 모델을 참조하여 각각의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 획득한다. 구체적으로, 정합 유닛(63)은 각각의 오디오 스트림을 원래의 성문 특징 모델의 입력값으로 개별적으로 사용하여 확률을 획득하거나, 각각의 오디오 스트림에 대응하는 정합도로서 참조하며, 원래의 성문 특징 모델과 정합하는 하나 이상의 오디오 스트림을 획득한다. 실제의 동작 프로세스에서, 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 성공적으로 정합된 오디오 스트림으로 선택하여, 그 획득된 오디오 스트림이 원래의 성문 특징 모델과 가장 관련 깊은 것으로 보장할 수 있으며, 이에 따라 성문 학습 말뭉치로서 사용되는 오디오 스트림은 순수하다. 모델 갱신 유닛(64)은 성공적으로 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 그런 다음 성문 등록을 수행하고 새로운 성문 특징 모델을 생성하거나, 정정된 성문 특징 모델로서 참조하며, 원래의 성문 특징 모델을 정정된 성문 특징 모델로 갱신한다. 최종적으로, 성문 특징 모델을 획득한 후, 상대적으로 높은 실용성의 전제 하에서 성문 특징 모델의 정확도를 높이는 목적이 달성되는 것을 보장한다.

[0099] **실시예 6:**

[0100] 도 7은 본 발명의 실시예 6에 따른 단말에 대한 구조도이다. 설명을 쉽게 하기 위해, 본 발명의 본 실시예와 관련된 부분만을 도시하고 있다. 본 발명의 실시예 6에서 제공하는 단말은 본 발명의 실시예 1 및 실시예 2의 방법을 실행하도록 구성될 수 있다. 설명을 쉽게 하기 위해, 본 발명의 본 실시예와 관련된 부분만을 도시하고 있다. 설명되지 않은 특정한 기술적 상세에 대해서는, 본 발명의 실시예 1 및 실시예 2를 참조하면 된다.

[0101] 단말은 샘플 획득 유닛(71), 원래의 모델 확립 유닛(72), 원래의 오디오 스트림 획득 유닛(73), 세그먼테이션 및 클러스너링 유닛(74), 정합 유닛(75), 모델 갱신 유닛(76)을 포함하며, 원래의 오디오 스트림 획득 유닛(73), 세그먼테이션 및 클러스너링 유닛(74), 정합 유닛(75), 및 모델 갱신 유닛(76)은 실시예 5에서의 원래의 오디오 스트림 획득 유닛(61), 세그먼테이션 및 클러스너링 유닛(62), 정합 유닛(63), 및 모델 갱신 유닛(64)에 각각 일대일 대응하고, 이에 대한 설명은 여기서 다시 설명하지 않는다.

[0102] 상기 샘플 획득 유닛(71)은 사전설정된 오디오 스트림 학습 샘플을 획득하고, 상기 사전설정된 오디오 스트림 학습 샘플을 상기 원래의 모델 확립 유닛(72)에 송신하도록 구성되어 있다.

[0103] 상기 원래의 모델 확립 유닛(72)은 상기 사전설정된 오디오 스트림 학습 샘플에 따라 상기 원래의 성문 특징 모델을 확립하도록 구성되어 있다.

[0104] 원래의 성문 특징 모델은 특정한 사람이나 여러 사람에 대한 성문 등록 프로세스 후에 형성되는 특징 모델이고, 이 등록 프로세스는 학습 말뭉치의 길이에 대한 요건을 가지지 않으며, 이를 오디오 스트림 학습 샘플이라고도 한다. 또한, 본 발명의 실시예에서 제공하는 방법은 정정된 모델에 대해 동적 정정을 지속적으로 수행할 수 있고, 원래의 성문 특징 모델은 기존의 방법을 사용하여 획득된 모델일 수 있으며, 본 발명의 본 실시예에서 제공하는 방법을 사용하여 정정된 모델일 수도 있다.

[0105] 본 발명의 본 실시예에서, 호출 듣기 상태로 들어간 후, 원래의 오디오 스트림 획득 유닛(73)은 듣기에 의해 오디오 스트림을 획득할 수 있으며, 여기서 오디오 스트림은 음성 레코더 또는 음성 채팅 소프트웨어를 사용하여 생성될 수 있다. 단말이 스마트폰인 경우를 예를 들어 사용한다.

[0106] 스마트폰이 호출 접속 상태에 있을 때, 사용자가 성문 학습 기능을 사용하는 것에 동의하는지를 질의하고, 사용자가 동의한 후, 호출에 참여하는 사용자 및 호출의 다른 상대방의 오디오 스트림이 기록될 수 있거나; 호출 동안 성문 학습 기능을 자동으로 가능하게 하는 스위치가 단말에 대해 구성되고, 사용자가 필요에 따라 스위치를 설정하거나; 성문 학습 기능이 단말에 대해 구성되고, 사용자는 오디오 스트림을 기록할 수 있다. 여러 사람이 전화 동안 또는 채팅 동안 차례로 대화에 참여할 수 있으므로, 이 경우에 획득된 원래의 오디오 스트림은 여러 사람의 오디오 데이터를 포함할 수 있다는 것에 유의해야 한다. 원래의 오디오 스트림 획득 유닛(73)에 의해 획득된 원래의 오디오 스트림은 화자의 다양한 억양, 말하는 속도, 및 감정에 대한 오디오 데이터를 망라

할 수 있으며, 모델 정확도에 대한 억양, 말하는 속도, 및 감정의 요인들의 효과를 감소시킨다. 또한, 사용자는 오디오 스트림을 획득하는 프로세스 동안의 특정한 횟수 및 지속시간을 가지는 오디오 스트림을 입력하지 않아도 되며, 이에 의해 사용자 동작의 복잡도를 감소시키고, 획득 프로세스에서의 실용성을 보장하며, 사용자 경험도 향상시킨다.

[0107] 도 7에 도시된 바와 같이, 세그먼테이션 및 클러스터링 유닛(74)은 구체적으로 세그먼테이션 유닛(741) 및 클러스터링 유닛(742)을 포함하며, 여기서

[0108] 상기 세그먼테이션 유닛(741)은 상기 원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하며, 여기서 상기 복수의 오디오 클립 중 각각의 오디오 클립은 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하며, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 상기 클러스터링 유닛(742)에 송신하도록 구성되어 있으며; 그리고

[0109] 상기 클러스터링 유닛(742)은 상기 세그먼테이션 유닛(741)에 의해 송신된, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 수신하고, 사전설정된 화자 클러스터링 알고리즘에 따라, 상기 적어도 한 명의 화자의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 상기 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성하도록 구성되어 있다.

[0110] 본 발명의 본 실시예에서, 세그먼테이션 유닛(741)은 원래의 오디오 스트림을 수 개의 오디오 클립으로 분할하고, 여기서 각각의 오디오 클립은 한 명의 화자의 오디오 정보만을 포함하며, 클러스터링 유닛(742)은 동일한 화자의 오디오 클립을 다시 클러스터링하여, 각자의 오디오 스트림을 생성한다. 최종적으로, 원래의 오디오 스트림은 상이한 화자를 나타내는 오디오 스트림으로 분할된다.

[0111] 도 7에 도시된 바와 같이, 정합 유닛(75)은 구체적으로 정합도 획득 유닛(751) 및 정합 오디오 스트림 획득 유닛(752)을 포함하며, 여기서

[0112] 상기 정합도 획득 유닛(751)은 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 상기 원래의 성문 특징 모델에 따라, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 획득하고, 상기 정합도를 상기 정합 오디오 스트림 획득 유닛(752)에 송신하도록 구성되어 있으며; 그리고

[0113] 상기 정합 오디오 스트림 획득 유닛(752)은 상기 정합도 획득 유닛(751)에 의해 송신된, 상기 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림과 원래의 성문 특징 모델 간의 정합도를 수신하고, 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 상기 성공적으로 정합된 오디오 스트림으로 선택하도록 구성되어 있다.

[0114] 본 발명의 본 실시예에서, 정합도 획득 유닛(7511)은 모든 오디오 스트림을 횡단하고, 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 원래의 성문 특징 모델에 따라 적어도 한 명의 화자 중 각각의 화자의 오디오 스트림 및 원래의 성문 특징 모델 간의 정합도를 획득한다. 구체적으로, 정합도 획득 유닛(7511)은 각각의 오디오 스트림을 원래의 성문 특징 모델의 입력값으로 개별적으로 사용하여 각각의 오디오 스트림에 대응하는 정합값을 획득하며, 상기 정합값은 구체적으로 성문 증명 알고리즘 인터페이스를 호출함으로써 획득될 수 있다. 그런 다음, 정합된 오디오 스트림 획득 유닛(752)은 원래의 성문 특징 모델과 정합하는 하나 이상의 오디오 스트림을 획득하고, 구체적으로 가장 높으면서 사전설정된 정합 임계값보다 큰 정합도에 대응하는 오디오 스트림을 성공적으로 정합된 오디오 스트림으로 선택하여, 그 획득된 오디오 스트림이 원래의 성문 특징 모델과 가장 관련 깊은 것으로 보장할 수 있으며, 이에 따라 성문 학습 말뭉치로서 사용되는 오디오 스트림은 순수하다.

[0115] 도 7에 도시된 바와 같이, 모델 갱신 유닛(76)은 구체적으로 정정 모델 획득 유닛(761) 및 모델 갱신 서브유닛(762)을 포함하며, 여기서

[0116] 상기 정정 모델 획득 유닛(761)은 상기 성공적으로 정합된 오디오 스트림 및 사전설정된 오디오 스트림 학습 샘플에 따라 정정된 성문 특징 모델을 생성하고, 상기 정정된 성문 특징 모델을 상기 모델 갱신 서브유닛(762)에 송신하도록 구성되어 있으며; 그리고

[0117] 상기 모델 갱신 서브유닛(762)은 상기 정정 모델 획득 유닛(761)에 의해 송신된 상기 정정된 성문 특징 모델을 수신하고, 상기 정정된 성문 특징 모델에 따라 상기 원래의 성문 특징 모델을 갱신하도록 구성되어 있다.

[0118] 본 발명의 본 실시예에서, 성공적으로 정합된 오디오 스트림은 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용된다. 즉, 원래의 성문 특징 모델을 생성하는 데 사용되는 오디오 스트림 학

습 샘플 및 성공적으로 정합된 오디오 스트림을 참조하여, 정정 모델 획득 유닛(761)은 성문 등록을 수행하고 새로운 성문 특징 모델을 생성하거나, 정정된 성문 특징 모델로서 참조하는 데 사용된다. 모델 갱신 서브유닛(762)은 원래의 성문 특징 모델을 정정된 성문 특징 모델로 갱신한다.

[0119] 본 발명의 본 실시예는 샘플 획득 유닛(71), 원래의 모델 확립 유닛(72), 원래의 오디오 스트림 획득 유닛(73), 세그먼테이션 및 클러스터링 유닛(74), 정합 유닛(75), 및 모델 갱신 유닛(76)을 포함하는 단말을 제공한다. 화자의 원래의 오디오 스트림 정보는 듣기에 의해 획득되며 성문 학습 말뭉치로서 사용되며, 원래의 오디오 스트림 정보는 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘을 사용함으로써 처리되어, 추가의 오디오 스트림 학습 샘플을 획득하며, 이에 따라 추가의 오디오 스트림 학습 샘플에 따라 원래의 성문 특징 모델에 대해 정정 및 갱신 동작이 수행되며, 이에 의해 상대적으로 높은 실용성의 전제 하에서 성문 특징 모델의 정확도를 높인다. 그러므로 성문 인식 정확도는 정정된 원래의 성문 특징 모델이 단말의 성문 잠금해제 솔루션에 적용될 때 현저하게 향상된다. 또한, 여러 사람의 스피치 오디오 스트림 학습 샘플에 대해 원래의 성문 특징 모델이 확립되면, 갱신된 원래의 성문 특징 모델은 여러 사람의 오디오 정보를 정확하게 인식하여 잠금해제 등을 수행하며, 이에 따라 잠금해제 프로세스가 더 지능적으로 된다.

[0120] 본 발명의 실시예에서 제공하는 성문 특징 모델을 갱신하는 방법에서, 적어도 한 명의 화자를 포함하는 원래의 오디오 스트림이 획득되며, 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림은 사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 획득되며, 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림은 원래의 성문 특징 모델과 개별적으로 정합되어, 성공적으로 정합된 오디오 스트림을 획득하며, 성공적으로 획득된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하며, 원래의 성문 특징 모델은 갱신된다. 이것은 기존의 방법을 사용하여 성문 특징 모델을 획득할 때, 성문 특징 모델의 정확도가 상대적으로 높은 실용성의 전제 하에서 향상되는 것을 보장할 수 없고, 그 결과 성문 특징 모델을 사용함으로써 인식 정확도가 향상될 수 없는 문제를 해결한다. 이것은 사용자 경험이 영향을 받지 않는 전자 하에서 성문 특징 모델의 정확도 및 인식 정확도를 향상시키며 상대적으로 높은 실용성이 보장된다.

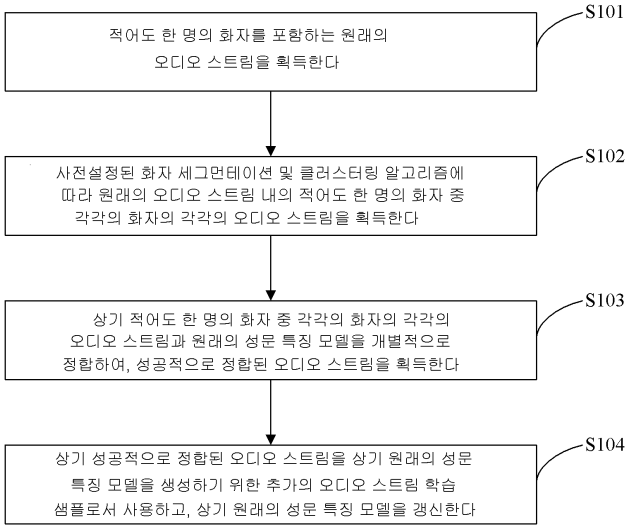
[0121] 본 명세서에서 설명된 실시예에서 설명된 예들을 조합함으로써, 유닛 및 알고리즘 단계들은 전자식 하드웨어, 컴퓨터 소프트웨어, 또는 이것들의 조합으로 실현될 수 있다는 것에 유의해야 한다. 하드웨어와 소프트웨어 간의 상호교환성을 명확하게 설명하기 위해, 위에서는 기능에 따라 각각의 예의 구성 및 단계를 개괄적으로 설명하였다. 이러한 기능들이 하드웨어 또는 소프트웨어로 수행되느냐는 것은 기술적 솔루션의 특별한 애플리케이션 및 설계 제약 조건에 달려 있다. 당업자라면 다양한 방법을 사용하여 각각의 특별한 애플리케이션에 대해 설명된 기능을 실행할 수 있을 것이며, 이것은 그 실행이 본 발명의 범주를 넘어서는 것으로 파악되어서는 안 된다.

[0122] 본 명세서에서 설명된 실시예를 조합하여, 방법 또는 알고리즘 단계는 하드웨어, 프로세서에 의해 실행되는 소프트웨어, 또는 이것들의 조합으로 실행될 수 있다. 소프트웨어 모듈은 랜덤 액세스 메모리(RAM), 리드-온리 메모리(ROM), 전기적으로 프로그래머블 ROM, 전기적으로 삭제 가능한 프로그머블 ROM, 레지스터, 하드디스크, 탈착식 디스크, CD-ROM, 또는 종래기술의 임의의 다른 형태의 저장 매체에 상주할 수 있다.

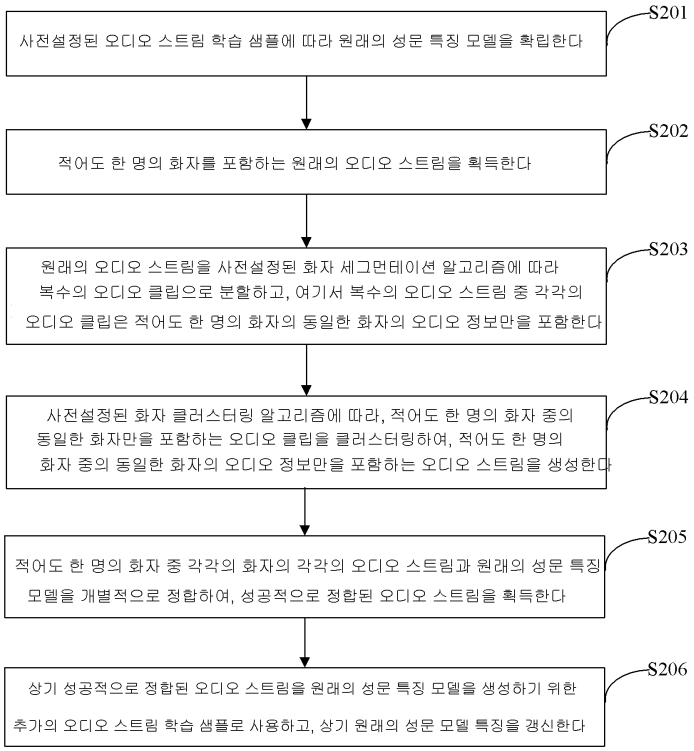
[0123] 본 발명의 목적, 기술적 솔루션 및 이로운 효과에 대해 특정한 실시예를 통해 상세히 설명하였다. 전술한 실시예는 단지 본 발명의 특정한 실행 모드에 지나지 않으며, 본 발명의 보호 범위를 제한하려는 것이 아님을 이해해야 한다. 본 발명의 정신 및 원리에 근거하여 이루어지는 모든 설정, 등가의 대체 및 개선은 본 발명의 보호 범위 내에 있게 된다.

## 도면

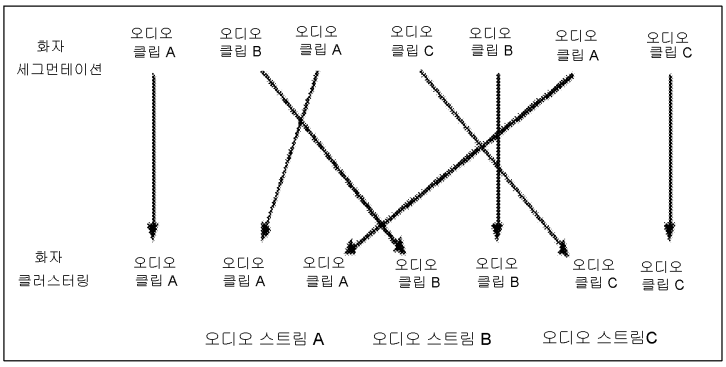
### 도면1


적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득한다
S101
사전설정된 화자 세그먼테이션 및 클러스터링 알고리즘에 따라 원래의 오디오 스트림 내의 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림을 획득한다
S102
상기 적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득한다
S103
상기 성공적으로 정합된 오디오 스트림을 상기 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로서 사용하고, 상기 원래의 성문 특징 모델을 갱신한다
S104

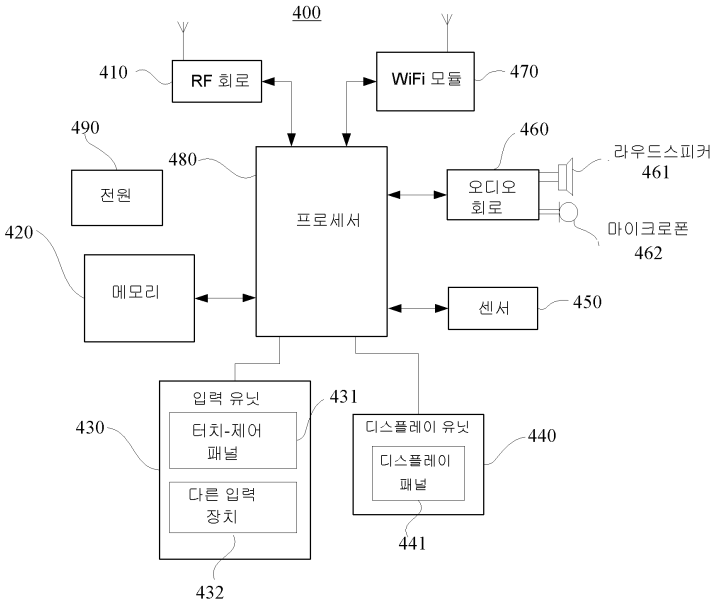
### 도면2


사전설정된 오디오 스트림 학습 샘플에 따라 원래의 성문 특징 모델을 확립한다
S201
적어도 한 명의 화자를 포함하는 원래의 오디오 스트림을 획득한다
S202
원래의 오디오 스트림을 사전설정된 화자 세그먼테이션 알고리즘에 따라 복수의 오디오 클립으로 분할하고, 여기서 복수의 오디오 스트림 중 각각의 오디오 클립은 적어도 한 명의 화자의 동일한 화자의 오디오 정보만을 포함한다
S203
사전설정된 화자 클러스터링 알고리즘에 따라, 적어도 한 명의 화자 중의 동일한 화자만을 포함하는 오디오 클립을 클러스터링하여, 적어도 한 명의 화자 중의 동일한 화자의 오디오 정보만을 포함하는 오디오 스트림을 생성한다
S204
적어도 한 명의 화자 중 각각의 화자의 각각의 오디오 스트림과 원래의 성문 특징 모델을 개별적으로 정합하여, 성공적으로 정합된 오디오 스트림을 획득한다
S205
상기 성공적으로 정합된 오디오 스트림을 원래의 성문 특징 모델을 생성하기 위한 추가의 오디오 스트림 학습 샘플로 사용하고, 상기 원래의 성문 모델 특징을 갱신한다
S206

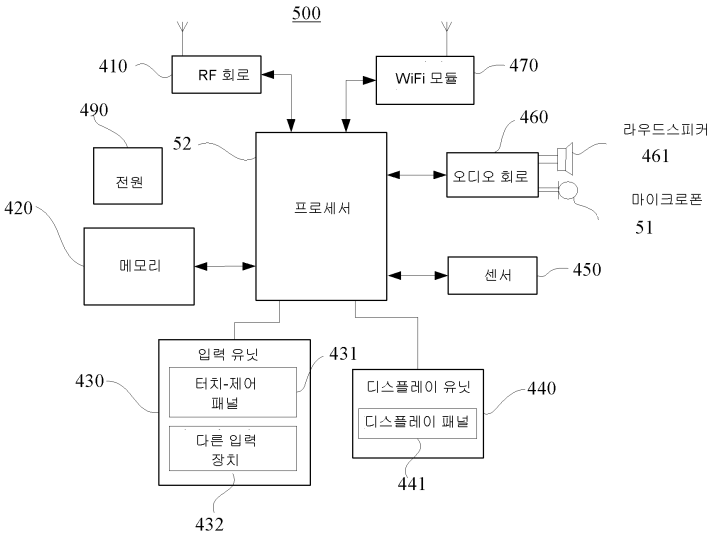
도면3


화자
세그먼테이션
오디오
클립 A
오디오
클립 B
오디오
클립 A
오디오
클립 C
오디오
클립 B
오디오
클립 A
오디오
클립 C
화자
클러스터링
오디오
클립 A
오디오
클립 A
오디오
클립 A
오디오
클립 B
오디오
클립 B
오디오
클립 C
오디오
클립 C
오디오 스트림 A
오디오 스트림 B
오디오 스트림C

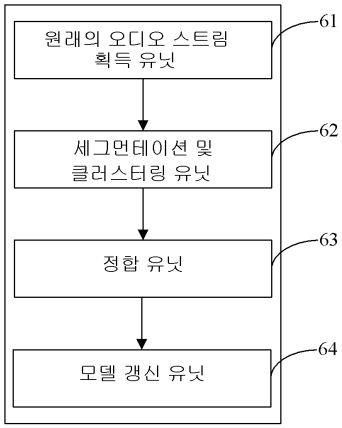
도면4


400
410
RF 회로
WiFi 모듈
470
490
전원
480
프로세서
460
오디오
회로
라우드스피커
461
마이크로폰
462
420
메모리
센서
450
입력 유닛
431
430
터치-제어
패널
다른 입력
장치
432
디스플레이 유닛
440
디스플레이
패널
441

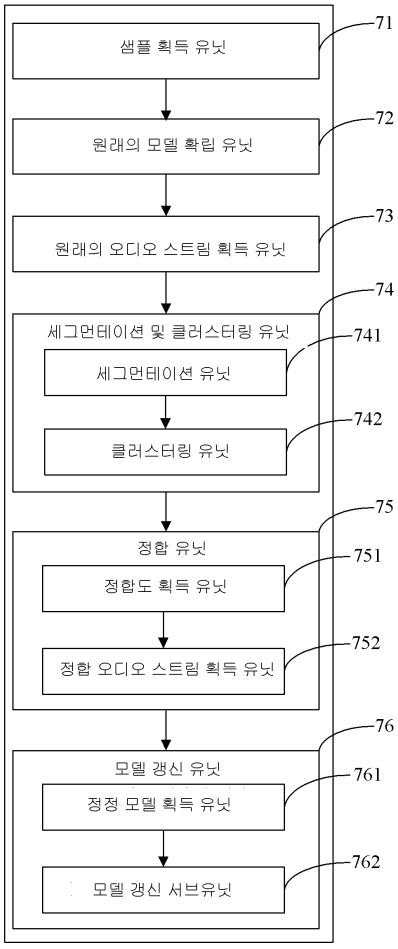
도면5


500
410
RF 회로
WiFi 모듈
470
490
52
460
라우드스피커
461
전원
프로세서
오디오 회로
마이크로폰
51
420
메모리
센서
450
입력 유닛
431
430
터치-제어 패널
디스플레이 유닛
440
다른 입력 장치
디스플레이 패널
441
432

도면6


원래의 오디오 스트림 획득 유닛
61
세그먼테이션 및 클러스터링 유닛
62
정합 유닛
63
모델 갱신 유닛
64

도면7


71
샘플 획득 유닛
72
원래의 모델 확립 유닛
73
원래의 오디오 스트림 획득 유닛
74
세그먼테이션 및 클러스터링 유닛
741
세그먼테이션 유닛
742
클러스터링 유닛
75
정합 유닛
751
정합도 획득 유닛
752
정합 오디오 스트림 획득 유닛
76
모델 갱신 유닛
761
정정 모델 획득 유닛
762
모델 갱신 서브유닛