(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0011307
(43) 공개일자 2023년01월20일

(51) 국제특허분류(Int. Cl.)
G10L 17/18 (2013.01) *G06F 21/32* (2013.01)
G06F 3/16 (2018.01) *G06N 20/00* (2019.01)
G10L 17/04 (2013.01) *G10L 17/24* (2013.01)
G16Y 20/20 (2020.01) *G16Y 30/10* (2020.01)
(52) CPC특허분류
G10L 17/18 (2013.01)
G06F 21/32 (2013.01)
(21) 출원번호 10-2022-7039892
(22) 출원일자(국제) 2021년04월15일
심사청구일자 없음
(85) 번역문제출일자 2022년11월15일
(86) 국제출원번호 PCT/US2021/027474
(87) 국제공개번호 WO 2021/211836
국제공개일자 2021년10월21일
(30) 우선권주장
63/010,504 2020년04월15일 미국(US)

(71) 출원인
핀드롭 시큐리티 인코포레이티드
미국 30308 조지아주 애틀란타 스위트 770 웨스트 피치트리 스트리트 엔더블유 817
(72) 발명자
코우리 엘리
미국 조지아 30308 아틀란타 스위트 770 엔더블유 웨스트 피치트리 스트리트 817
시바라만 가네쉬
미국 조지아 30308 아틀란타 스위트 770 엔더블유 웨스트 피치트리 스트리트 817
(뒷면에 계속)
(74) 대리인
신성특허법인(유한)

전체 청구항 수 : 총 110 항

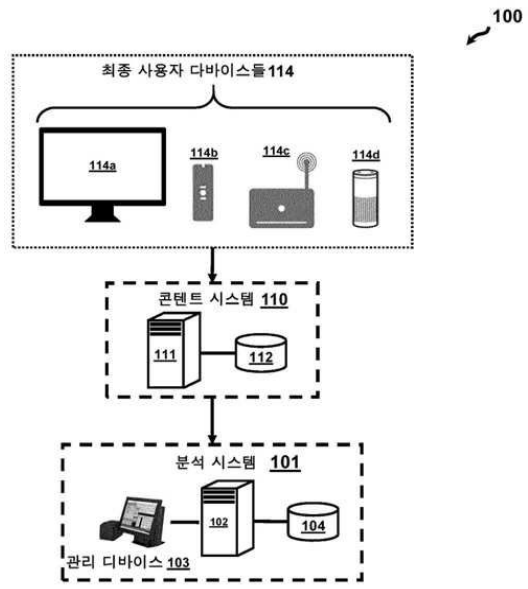
(54) 발명의 명칭 **수동적 및 지속적 다중-화자 음성 생체 인식**

(57) 요 약

본원에 설명된 실시예들은 수동적, 능동적, 지속적 또는 정적 동작들, 또는 이들의 조합이 가능한 기계 학습 아키텍처들을 실행하는 음성 생체 인식 시스템을 제공한다. 시스템들은 수동적으로 및/또는 지속적으로, 일부 경우에는 능동적으로 및/또는 정적으로 추가하여, 화자들이 에지 디바이스(예를 들어, 자동차, 텔레비전, 라디오,

(뒷면에 계속)

대 표 도 - 도1



전화) 내부 또는 주변에서 말할 때 화자들을 등록한다. 시스템은 재구성 동작들을 위해 새 화자가 프롬프트된 발화들을 미러링할 필요 없이 즉석에서 사용자들을 식별한다. 시스템은 화자들이 시스템에 발화들을 제공할 때 화자 프로파일들 관리한다. 기계 학습 아키텍처들은 가능한 한 화자 ID들에 대한 지식 없이 수동적 및 지속적 음성 생체 인식 시스템을 구현한다. 시스템은 비지도 방식으로 ID들을 생성하며, 때로는 수동적으로 알려지거나 알려지지 않은 화자들을 등록하고 인식한다. 시스템은 오버-더-톱 서비스들 및 IoT 디바이스들(예를 들어, 개인 어시스턴트들, 차량들), 및 콜 센터들을 위한 미디어 콘텐트를 포함하는, 다양한 애플리케이션들에 걸쳐 개인화 및 보안을 제공한다.

(52) CPC특허분류
G06F 3/167 (2013.01)
G06N 20/00 (2021.08)
G10L 17/04 (2013.01)
G10L 17/24 (2013.01)
G16Y 20/20 (2020.01)
G16Y 30/10 (2020.01)

(72) 발명자
쿠마르 아브로쉬
미국 조지아 30308 아틀란타 스위트 770 엔더블유 웨스트 피치트리 스트리트 817

앤톨릭-소번 이반
미국 조지아 30308 아틀란타 스위트 770 엔더블유 웨스트 피치트리 스트리트 817

명 세 서

청구범위

청구항 1

컴퓨터 구현 방법에 있어서,

상기 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩(inbound embedding)을 추출하는 단계;

상기 컴퓨터에 의해, 화자 프로파일 데이터베이스의 화자 프로파일에 저장된 성문(voiceprint)과 상기 인바운드 임베딩 사이의 거리에 기초하여 유사도 점수(similarity score)를 생성하는 단계; 및

상기 컴퓨터가 상기 인바운드 임베딩에 대한 상기 유사도 점수가 유사도 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

상기 컴퓨터에 의해, 상기 화자 프로파일 데이터베이스에, 상기 인바운드 임베딩을 포함하는 상기 인바운드 화자에 대한 새로운 화자 프로파일을 생성하는 단계로서, 상기 새로운 화자 프로파일은 상기 인바운드 임베딩을 새로운 성문으로서 저장하는 데이터베이스 레코드인, 상기 생성하는 단계를 포함하는, 방법.

청구항 2

제1항에 있어서,

상기 컴퓨터에 의해, 중개 서버(intermediary server)를 통해 최종 사용자 디바이스로부터 상기 인바운드 오디오 신호를 수신하는 단계; 및

상기 컴퓨터에 의해, 상기 새로운 화자 프로파일과 연관된 새로운 화자 식별자를 상기 서버에 전송하는 단계를 더 포함하는, 방법.

청구항 3

제2항에 있어서,

상기 최종 사용자 디바이스는 스마트 텔레비전, 텔레비전에 결합된 미디어 디바이스, 및 에지(edge) 디바이스 중 적어도 하나인, 방법.

청구항 4

제1항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하는 단계를 더 포함하며, 상기 컴퓨터가 상기 인바운드 오디오 신호로부터 추출한 상기 하나 이상의 특징들에 상기 기계 학습 모델을 적용함으로써 상기 컴퓨터가 상기 인바운드 임베딩을 생성하는, 방법.

청구항 5

제1항에 있어서,

상기 컴퓨터에 의해, 상기 제2 인바운드 오디오 신호에 상기 기계 학습 모델을 적용함으로써 제2 인바운드 신호로부터 제2 인바운드 임베딩을 추출하는 단계;

상기 컴퓨터에 의해, 상기 새로운 화자 프로파일에 저장된 상기 새로운 성문과 상기 제2 인바운드 임베딩 사이의 상기 거리에 기초하여 제2 유사도 점수를 생성하는 단계; 및

상기 컴퓨터가 상기 제2 인바운드 임베딩에 대한 상기 제2 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여,

상기 컴퓨터에 의해, 상기 제2 인바운드 신호에 기초하여 상기 인바운드 화자에 대한 상기 새로운 성문을 업데

이트하는 단계를 더 포함하는, 방법.

청구항 6

제1항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호와 연관된 가입자 식별자를 수신하는 단계; 및

상기 컴퓨터에 의해, 상기 화자 프로파일 데이터베이스에 저장된 상기 가입자 식별자와 연관된 하나 이상의 화자 프로파일들을 식별하는 단계로서, 상기 컴퓨터는 상기 가입자 식별자와 연관된 상기 하나 이상의 화자 프로파일들에 저장된 하나 이상의 성문들에 기초하여 상기 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하는, 상기 식별하는 단계를 더 포함하는, 방법.

청구항 7

제6항에 있어서,

상기 가입자 식별자는 하나 이상의 화자 식별자들과 연관되고, 각 화자 프로파일은 대응하는 화자 식별자와 연관되는, 방법.

청구항 8

제7항에 있어서,

상기 가입자 식별자 및 각 화자 식별자 중 적어도 하나는 익명(anonymized) 식별자인, 방법.

청구항 9

제1항에 있어서,

상기 컴퓨터는 하나 이상의 인바운드 임베딩들에 기초하여 상기 새로운 성문을 생성하고, 상기 방법은,

상기 컴퓨터에 의해, 상기 하나 이상의 인바운드 임베딩들에 기초하여 상기 새로운 성문에 대한 하나 이상의 성숙(maturity) 인자들을 식별하는 단계; 및

상기 컴퓨터에 의해, 상기 하나 이상의 성숙 인자들에 기초하여 상기 새로운 성문에 대한 성숙 레벨을 결정하는 단계를 더 포함하는, 방법.

청구항 10

제9항에 있어서,

상기 컴퓨터에 의해, 상기 컴퓨터가 상기 새로운 성문에 대한 상기 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 상기 새로운 화자 프로파일의 새로운 유사도 임계값을 업데이트하는 단계를 더 포함하는, 방법.

청구항 11

제9항에 있어서,

상기 컴퓨터가 상기 성숙 레벨이 성숙 임계값에 못 미친다고 결정하는 것에 응답하여,

상기 컴퓨터에 의해, 능동적 등록 프롬프트들을 생성하는 단계로서, 상기 능동적 등록 프롬프트는 추가 인바운드 오디오 신호에 대한 요청을 디스플레이하도록 구성된 사용자 인터페이스를 포함하는, 상기 생성하는 단계;

상기 컴퓨터에 의해, 상기 추가 인바운드 신호로부터 추가 임베딩을 추출하는 단계; 및

상기 컴퓨터에 의해, 상기 추가 인바운드 신호로부터 추출된 추가 임베딩에 따라 상기 새로운 성문을 업데이트하는 단계를 더 포함하는, 방법.

청구항 12

제9항에 있어서,

상기 컴퓨터에 의해, 상기 컴퓨터가 상기 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 임시 프로파일로부터 영구 프로파일로 상기 새로운 화자 프로파일을 업데이트하는 단계를 더 포함하는, 방법.

청구항 13

시스템에 있어서,

화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및

컴퓨터로서,

인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하고;

상기 화자 프로파일 데이터베이스의 화자 프로파일에 저장된 성문과 상기 인바운드 임베딩 사이의 거리에 기초하여 유사도 점수를 생성하고;

상기 컴퓨터가 상기 인바운드 임베딩에 대한 상기 유사도 점수가 유사도 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

상기 화자 프로파일 데이터베이스에, 상기 인바운드 임베딩을 포함하는 상기 인바운드 화자에 대한 새로운 화자 프로파일을 생성하도록 구성되되, 상기 새로운 화자 프로파일은 상기 인바운드 임베딩을 새로운 성문으로 저장하는 데이터베이스 레코드인, 프로세서를 포함하는, 상기 컴퓨터를 포함하는, 시스템.

청구항 14

제13항에 있어서,

상기 컴퓨터는,

중개 서버를 통해 최종 사용자 디바이스로부터 상기 인바운드 오디오 신호를 수신하고;

상기 새로운 화자 프로파일과 연관된 새로운 화자 식별자를 상기 서버에 전송하도록 더 구성되는, 시스템.

청구항 15

제14항에 있어서,

상기 최종 사용자 디바이스는 스마트 텔레비전, 텔레비전에 결합된 미디어 디바이스 및 에지 디바이스 중 적어도 하나인, 시스템.

청구항 16

제13항에 있어서,

상기 컴퓨터는 상기 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하도록 더 구성되며, 상기 컴퓨터가 상기 인바운드 오디오 신호로부터 추출한 상기 하나 이상의 특징들에 상기 기계 학습 모델을 적용함으로써 상기 컴퓨터가 상기 인바운드 임베딩을 생성하는, 시스템.

청구항 17

제13항에 있어서,

상기 컴퓨터는,

상기 제2 인바운드 오디오 신호에 상기 기계 학습 모델을 적용함으로써 제2 인바운드 신호로부터 제2 인바운드 임베딩을 추출하고;

상기 새로운 화자 프로파일에 저장된 상기 새로운 성문과 상기 제2 인바운드 임베딩 사이의 상기 거리에 기초하여 제2 유사도 점수를 생성하고;

상기 컴퓨터가 상기 제2 인바운드 임베딩에 대한 상기 제2 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여,

상기 제2 인바운드 신호에 기초하여 상기 인바운드 화자에 대한 상기 새로운 성문을 업데이트하도록 더 구성되는, 시스템.

청구항 18

제13항에 있어서,

상기 컴퓨터는,

상기 인바운드 오디오 신호와 연관된 가입자 식별자를 수신하고;

상기 화자 프로파일 데이터베이스에 저장된 상기 가입자 식별자와 연관된 하나 이상의 화자 프로파일들을 식별하되, 상기 컴퓨터는 상기 가입자 식별자와 연관된 상기 하나 이상의 화자 프로파일들에 저장된 하나 이상의 성문들에 기초하여 상기 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하도록 더 구성되는, 시스템.

청구항 19

제18항에 있어서,

상기 가입자 식별자는 하나 이상의 화자 식별자들과 연관되고, 각 화자 프로파일은 대응하는 화자 식별자와 연관되는, 시스템.

청구항 20

제19항에 있어서,

상기 가입자 식별자 및 각 화자 식별자 중 적어도 하나는 익명 식별자인, 시스템.

청구항 21

제13항에 있어서,

상기 컴퓨터는 하나 이상의 인바운드 임베딩들에 기초하여 상기 새로운 성문을 생성하고, 상기 컴퓨터는,

상기 하나 이상의 인바운드 임베딩들에 기초하여 상기 새로운 성문에 대한 하나 이상의 성숙 인자들을 식별하고;

상기 하나 이상의 성숙 인자들에 기초하여 상기 새로운 성문에 대한 성숙 레벨을 결정하도록 더 구성되는, 시스템.

청구항 22

제21항에 있어서,

상기 컴퓨터가 상기 새로운 성문에 대한 상기 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 상기 새로운 화자 프로파일의 새로운 유사도 임계값을 업데이트하도록 더 구성되는, 시스템.

청구항 23

제21항에 있어서,

상기 컴퓨터는,

상기 성숙 레벨이 성숙 임계값에 못 미친다고 결정하는 것에 응답하여,

능동적 등록 프롬프트들을 생성하되, 상기 능동적 등록 프롬프트는 추가 인바운드 오디오 신호에 대한 요청을 디스플레이하도록 구성된 사용자 인터페이스를 포함하고;

상기 추가 인바운드 신호로부터 추가 임베딩을 추출하고;

상기 추가 인바운드 신호로부터 추출된 추가 임베딩에 따라 상기 새로운 성문을 업데이트하도록 더 구성되는, 시스템.

청구항 24

제21항에 있어서,

상기 컴퓨터가 상기 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 임시 프로파일로부터 영구 프로파일로 상기 새로운 화자 프로파일을 업데이트하도록 더 구성되는, 시스템.

청구항 25

컴퓨터 구현 방법에 있어서,

컴퓨터에 의해, 복수의 인바운드 화자들의 복수의 발화(utterance)들을 포함하는 인바운드 오디오 신호를 수신하는 단계;

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호에 기계 학습 아키텍처를 적용하여 상기 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들을 추출하는 단계;

상기 복수의 인바운드 화자들의 각 인바운드 화자에 대해,

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 인바운드 임베딩에 기초하여 하나 이상의 유사도 점수들을 생성하는 단계로서, 상기 인바운드 화자에 대한 각 유사도 점수는 상기 인바운드 임베딩과 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들 사이의 거리를 나타내는, 상기 생성하는 단계; 및

상기 컴퓨터에 의해, 하나 이상의 성문들로부터 상기 인바운드 화자에 대한 가장 가까운 성문을 식별하는 단계로서, 상기 가장 가까운 성문은 상기 인바운드 화자에 대해 생성된 상기 하나 이상의 유사도 점수들 중 최대 유사도 점수에 대응하는, 상기 식별하는 단계; 및

각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족하는 경우,

상기 컴퓨터에 의해, 상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 상기 인바운드 화자에 대한 상기 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하는 단계를 포함하는, 방법.

청구항 26

제25항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호와 함께 수신된 가입자 식별자에 기초하여 상기 화자 프로파일 데이터베이스에 저장된 상기 하나 이상의 성문들을 식별하는 단계를 더 포함하는, 방법.

청구항 27

제25항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 화자 임베딩에 대한 상기 최대 유사도 점수가 상기 하나 이상의 유사도 점수들을 충족한다고 결정하는 단계; 및

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 상기 가장 가까운 성문을 포함하는 상기 화자 데이터베이스의 화자 프로파일을 식별하는 단계로서, 상기 화자 프로파일은 화자 식별자를 포함하는, 상기 식별하는 단계를 더 포함하는, 방법.

청구항 28

제25항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 상기 최대 유사도 점수가 제1 유사도 임계값을 충족하고 제2 유사도 임계값에 못 미친다고 결정하는 단계로서, 상기 제1 유사도 임계값은 제2 유사도 임계값보다 비교적 낮은, 상기 결정하는 단계; 및

상기 컴퓨터에 의해, 상기 인바운드 화자의 상기 인바운드 임베딩을 포함하도록 상기 화자 데이터베이스에 저장된 약한 임베딩 목록을 업데이트하는 단계를 더 포함하는, 방법.

청구항 29

제28항에 있어서,

상기 컴퓨터에 의해, 상기 가입자 식별자와 연관된 상기 화자 데이터베이스의 하나 이상의 화자 프로파일들에 대한 재클러스터링 동작을 실행하는 단계;

상기 컴퓨터에 의해, 상기 재클러스터링 동작에 기초하여 상기 약한 임베딩 목록의 상기 인바운드 임베딩에 대한 상기 최대 유사도 점수를 업데이트하는 단계; 및

상기 약한 임베딩 목록의 상기 인바운드 임베딩에 대한 상기 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족한다고 결정하는 것에 응답하여,

상기 컴퓨터에 의해, 상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 상기 인바운드 화자에 대한 상기 인바운드 임베딩을 포함하도록 상기 화자 프로파일 데이터베이스를 업데이트하는 단계를 더 포함하는, 방법.

청구항 30

제25항에 있어서,

상기 컴퓨터에 의해, 상기 가입자 식별자에 대한 재클러스터링 동작을 수행하기 위한 트리거링 조건을 검출하는 단계; 및

상기 컴퓨터에 의해, 상기 가입자 식별자와 연관된 하나 이상의 화자 프로파일들 중 상기 복수의 성문들에 대한 계층적 클러스터링 동작을 실행하는 단계를 더 포함하는, 방법.

청구항 31

제30항에 있어서,

상기 컴퓨터에 의해, 기존 화자 프로파일로부터 새로운 성문 클러스터와 연관된 상기 화자 식별자를 업데이트하는 단계로서, 상기 새로운 성문 클러스터는 상기 가입자 식별자와 연관된 상기 하나 이상의 화자 프로파일들에 상기 계층적 클러스터링 동작을 적용함으로써 생성되는, 상기 업데이트하는 단계를 더 포함하는, 방법.

청구항 32

제25항에 있어서,

상기 컴퓨터에 의해, 중개 서버를 통해 최종 사용자 디바이스로부터 상기 인바운드 오디오 신호를 수신하는 단계; 및

상기 컴퓨터에 의해, 상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 상기 화자 식별자를 전송하는 단계를 더 포함하는, 방법.

청구항 33

제25항에 있어서,

컴퓨터에 의해, 복수의 인바운드 화자들의 복수의 발화들을 포함하는 상기 인바운드 오디오 신호를 수신하는 단계를 더 포함하는, 방법.

청구항 34

제25항에 있어서,

상기 컴퓨터에 의해, 상기 하나 이상의 특징들에 상기 기계 학습 아키텍처를 적용하여 하나 이상의 오디오 이벤트들을 식별하여 환경 설정을 검출하는 단계; 및

상기 컴퓨터에 의해, 상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 상기 환경 설정 및 각 화자 식별자의 표시자를 중개 서버에 전송하는 단계를 더 포함하는, 방법.

청구항 35

제25항에 있어서,

각 최대 유사도 점수가 상기 하나 이상의 유사도 점수 임계값들에 못 미치는 경우,

상기 컴퓨터에 의해, 상기 인바운드 임베딩을 포함하는 상기 가입자 식별자와 연관된 새로운 화자 프로파일을 생성하는 단계; 및

상기 컴퓨터에 의해, 상기 새로운 화자 프로파일 및 상기 인바운드 화자에 대한 상기 인바운드 임베딩을 포함하도록 상기 화자 프로파일 데이터베이스를 업데이트하는 단계를 더 포함하는, 방법.

청구항 36

시스템에 있어서,

화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및

컴퓨터로서,

복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하고;

상기 인바운드 오디오 신호에 기계 학습 아키텍처를 적용하여 상기 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들을 추출하고;

상기 복수의 인바운드 화자들의 각 인바운드 화자에 대해,

상기 인바운드 화자에 대한 인바운드 임베딩에 기초하여 하나 이상의 유사도 점수들을 생성하되, 상기 인바운드 화자에 대한 각 유사도 점수는 상기 인바운드 임베딩과 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들 사이의 거리를 나타내고;

상기 하나 이상의 성문들로부터 상기 인바운드 화자에 대한 가장 가까운 성문을 식별하되, 상기 가장 가까운 성문은 상기 인바운드 화자에 대해 생성된 상기 하나 이상의 유사도 점수들 중 최대 유사도 점수에 대응하고;

각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족하는 경우,

상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 상기 인바운드 화자에 대한 상기 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함하는, 시스템.

청구항 37

제36항에 있어서,

상기 컴퓨터는 상기 인바운드 오디오 신호와 함께 수신된 가입자 식별자에 기초하여 상기 화자 프로파일 데이터베이스에 저장된 상기 하나 이상의 성문들을 식별하도록 더 구성되는, 시스템.

청구항 38

제36항에 있어서,

상기 컴퓨터는,

상기 인바운드 화자 임베딩에 대한 상기 최대 유사도 점수가 상기 하나 이상의 유사도 점수들을 충족한다고 결정하고;

상기 인바운드 화자에 대한 상기 가장 가까운 성문을 포함하는 상기 화자 데이터베이스의 화자 프로파일을 식별하되, 상기 화자 프로파일은 화자 식별자를 포함하도록 더 구성되는, 시스템.

청구항 39

제36항에 있어서,

상기 컴퓨터는,

상기 인바운드 화자에 대한 상기 최대 유사도 점수가 제1 유사도 임계값을 충족하고 제2 유사도 임계값에 못 미

친다고 결정하되, 상기 제1 유사도 임계값은 제2 유사도 임계값보다 비교적 낮고;

상기 인바운드 화자의 상기 인바운드 임베딩을 포함하도록 상기 화자 데이터베이스에 저장된 약한 임베딩 목록을 업데이트하도록 더 구성되는, 시스템.

청구항 40

제39항에 있어서,

상기 컴퓨터는,

상기 가입자 식별자와 연관된 상기 화자 데이터베이스의 하나 이상의 화자 프로파일들에 대한 재클러스터링 동작을 실행하고;

상기 재클러스터링 동작에 기초하여 상기 약한 임베딩 목록의 상기 인바운드 임베딩에 대한 상기 최대 유사도 점수를 업데이트하고;

상기 약한 임베딩 목록의 상기 인바운드 임베딩에 대한 상기 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족한다고 결정하는 것에 응답하여,

상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 상기 인바운드 화자에 대한 상기 인바운드 임베딩을 포함하도록 상기 화자 프로파일 데이터베이스를 업데이트하도록 더 구성되는, 시스템.

청구항 41

제36항에 있어서,

상기 컴퓨터는,

상기 가입자 식별자에 대한 재클러스터링 동작을 수행하기 위한 트리거링 조건을 검출하고;

상기 가입자 식별자와 연관된 하나 이상의 화자 프로파일들 중 상기 복수의 성문들에 대한 계층적 클러스터링 동작을 실행하도록 더 구성되는, 시스템.

청구항 42

제41항에 있어서,

상기 컴퓨터는 기존 화자 프로파일로부터 새로운 성문 클러스터와 연관된 상기 화자 식별자를 업데이트하되, 상기 새로운 성문 클러스터는 상기 가입자 식별자와 연관된 상기 하나 이상의 화자 프로파일들에 상기 계층적 클러스터링 동작을 적용함으로써 생성되도록 더 구성되는, 시스템.

청구항 43

제36항에 있어서,

상기 컴퓨터는,

중개 서버를 통해 최종 사용자 디바이스로부터 상기 인바운드 오디오 신호를 수신하고;

상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 상기 화자 식별자를 전송하도록 더 구성되는, 시스템.

청구항 44

제36항에 있어서,

상기 컴퓨터가 복수의 인바운드 화자들의 복수의 발화들을 포함하는 상기 인바운드 오디오 신호를 수신하도록 더 구성되는, 시스템.

청구항 45

제36항에 있어서,

상기 컴퓨터는,

상기 하나 이상의 특징들에 상기 기계 학습 아키텍처를 적용하여 하나 이상의 오디오 이벤트들을 식별하여 환경 설정을 검출하고;

상기 하나 이상의 유사도 점수 임계값들을 충족하는 상기 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 상기 환경 설정 및 각 화자 식별자의 표시자를 중개 서버에 전송하도록 더 구성되는, 시스템.

청구항 46

제36항에 있어서,

상기 컴퓨터는,

각 최대 유사도 점수가 상기 하나 이상의 유사도 점수 임계값들에 못 미치는 경우,

상기 인바운드 임베딩을 포함하는 상기 가입자 식별자와 연관된 새로운 화자 프로파일을 생성하고;

상기 새로운 화자 프로파일 및 상기 인바운드 화자에 대한 상기 인바운드 임베딩을 포함하도록 상기 화자 프로파일 데이터베이스를 업데이트하도록 더 구성되는, 시스템.

청구항 47

컴퓨터 구현 방법에 있어서,

컴퓨터에 의해, 콘텐트 서버를 통해 최종 사용자 디바이스로부터 인바운드 화자에 대한 인바운드 오디오 신호를 수신하는 단계;

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호에 기계 학습 모델을 적용하여 상기 인바운드 화자에 대한 인바운드 임베딩을 추출하는 단계;

상기 컴퓨터에 의해, 상기 인바운드 임베딩과 화자 데이터베이스의 화자 프로파일에 저장된 성문 사이의 거리에 기초하여 상기 인바운드 임베딩에 대한 유사도 점수를 생성하는 단계로서, 상기 유사도 점수는 하나 이상의 유사도 점수 임계값들을 충족하는, 상기 생성하는 단계;

상기 컴퓨터에 의해, 상기 화자 프로파일에서 상기 콘텐트 서버에 대한 하나 이상의 콘텐트 특성들에 대응하는 하나 이상의 화자 특성들을 식별하는 단계; 및

상기 컴퓨터에 의해, 상기 인바운드 화자와 연관된 상기 하나 이상의 화자 특성들을 상기 미디어 콘텐트 서버에 전송하는 단계를 포함하는, 방법.

청구항 48

제47항에 있어서,

상기 컴퓨터에 의해, 상기 콘텐트 서버로부터 수신된 가입자 식별자에 기초하여 상기 화자 프로파일을 식별하는 단계를 더 포함하는, 방법.

청구항 49

제48항에 있어서,

상기 컴퓨터에 의해 수신된 상기 가입자 식별자는 제1 익명 식별자이고, 상기 가입자 식별자는 하나 이상의 화자들에 대한 하나 이상의 화자 식별자들과 연관되고, 각 화자 식별자는 각 화자 프로파일에 대응하는 제2 익명 식별자인, 방법.

청구항 50

제47항에 있어서,

상기 하나 이상의 화자 특성들은 연령 특성 및 성별 특성 중 적어도 하나를 포함하는, 방법.

청구항 51

제47항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호에 제2 기계 학습 모델을 적용함으로써 상기 인바운드 화자의 연령 특성 및 성별 특성 중 적어도 하나를 결정하는 단계를 더 포함하는, 방법.

청구항 52

제47항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 상기 화자 프로파일에 저장된 연령 특성에 기초하여 상기 인바운드 화자의 연령을 결정하는 단계를 더 포함하는, 방법.

청구항 53

제47항에 있어서,

상기 컴퓨터에 의해, 상기 콘텐트 서버로부터 상기 인바운드 화자에 대한 적어도 하나의 화자 특성을 나타내는 화자 정보를 수신하는 단계;

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 상기 화자 프로파일에 상기 적어도 하나의 화자 특성을 저장하는 단계; 및

상기 컴퓨터에 의해, 상기 화자 프로파일의 상기 적어도 하나의 화자 특성에 부분적으로 기초하여 상기 화자 프로파일을 식별하는 단계를 더 포함하는, 방법.

청구항 54

제47항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호와 함께 콘텐트 서버로부터 인바운드 인증 데이터를 수신하는 단계; 및

상기 컴퓨터에 의해, 상기 유사도 점수가 상기 유사도 임계값을 충족하고 상기 인바운드 인증 데이터가 상기 화자 프로파일에 저장된 예상된 인증 데이터를 충족하는 것에 기초하여 상기 인바운드 화자를 인증하는 단계를 더 포함하는, 방법.

청구항 55

제54항에 있어서,

상기 인증 데이터는 상기 최종 사용자 디바이스, 화자 정보 및 생체 인증 정보와 연관된 최종 사용자 디바이스 정보, 메타데이터 중 적어도 하나를 포함하는, 방법.

청구항 56

제47항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하는 단계; 및

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 상기 하나 이상의 특징들에 기초하여 스푸핑(spoofing) 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하는 단계를 더 포함하는, 방법.

청구항 57

제47항에 있어서,

상기 컴퓨터에 의해, 미디어 콘텐트 데이터베이스에서 상기 하나 이상의 화자 특성들에 대응하는 하나 이상의 콘텐트 특성들을 갖는 하나 이상의 미디어 콘텐트 파일들을 식별하는 단계를 더 포함하는, 방법.

청구항 58

제57항에 있어서,

상기 컴퓨터에 의해, 상기 화자의 상기 하나 이상의 화자 특성들 중 연령 특성이 연령 제한 콘텐트의 상기 하나 이상의 콘텐트 특성들의 해당 연령 제한 특성을 충족하는 것에 기초하여 상기 콘텐트 데이터베이스에서 상기 연령 제한 콘텐트를 활성화하는 단계를 더 포함하는, 방법.

청구항 59

시스템에 있어서,

복수의 화자 프로파일들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및

서버로서,

콘텐트 서버를 통해 최종 사용자 디바이스로부터 인바운드 화자에 대한 인바운드 오디오 신호를 수신하고;

상기 인바운드 오디오 신호에 기계 학습 모델을 적용하여 상기 인바운드 화자에 대한 인바운드 임베딩을 추출하고;

상기 인바운드 임베딩과 상기 화자 데이터베이스의 화자 프로파일에 저장된 성문 사이의 거리에 기초하여 상기 인바운드 임베딩에 대한 유사도 점수를 생성하되, 상기 유사도 점수는 하나 이상의 유사도 점수 임계값들을 충족하고;

상기 화자 프로파일에서 상기 콘텐트 서버에 대한 하나 이상의 콘텐트 특성들에 대응하는 하나 이상의 화자 특성들을 식별하고;

상기 인바운드 화자와 연관된 상기 하나 이상의 화자 특성들을 상기 미디어 콘텐트 서버에 전송하도록 구성된 프로세서를 포함하는, 상기 서버를 포함하는, 시스템.

청구항 60

제59항에 있어서,

상기 컴퓨터는 상기 콘텐트 서버로부터 수신된 가입자 식별자에 기초하여 상기 화자 프로파일을 식별하도록 더 구성되는, 시스템.

청구항 61

제60항에 있어서,

상기 컴퓨터에 의해 수신된 상기 가입자 식별자는 제1 익명 식별자이고, 상기 가입자 식별자는 하나 이상의 화자들에 대한 하나 이상의 화자 식별자들과 연관되고, 각 화자 식별자는 각 화자 프로파일에 대응하는 제2 익명 식별자인, 시스템.

청구항 62

제59항에 있어서,

상기 하나 이상의 화자 특성들은 연령 특성 및 성별 특성 중 적어도 하나를 포함하는, 시스템.

청구항 63

제59항에 있어서,

상기 컴퓨터는 상기 인바운드 오디오 신호에 제2 기계 학습 모델을 적용함으로써 상기 인바운드 화자의 연령 특성 및 성별 특성 중 적어도 하나를 결정하도록 더 구성되는, 시스템.

청구항 64

제59항에 있어서,

상기 컴퓨터는 상기 인바운드 화자에 대한 상기 화자 프로파일에 저장된 연령 특성에 기초하여 상기 인바운드 화자의 연령을 결정하도록 더 구성되는, 시스템.

청구항 65

제59항에 있어서,

상기 컴퓨터는,

상기 콘텐트 서버로부터 상기 인바운드 화자에 대한 적어도 하나의 화자 특성을 나타내는 화자 정보를 수신하고;

상기 인바운드 화자에 대한 상기 화자 프로파일에 상기 적어도 하나의 화자 특성을 저장하고;

상기 화자 프로파일의 상기 적어도 하나의 화자 특성에 부분적으로 기초하여 상기 화자 프로파일을 식별하도록 더 구성되는, 시스템.

청구항 66

제59항에 있어서,

상기 컴퓨터는,

상기 인바운드 오디오 신호와 함께 콘텐트 서버로부터 인바운드 인증 데이터를 수신하고;

상기 유사도 점수가 상기 유사도 임계값을 충족하고 상기 인바운드 인증 데이터가 상기 화자 프로파일에 저장된 예상된 인증 데이터를 충족하는 것에 기초하여 상기 인바운드 화자를 인증하도록 더 구성되는, 시스템.

청구항 67

제66항에 있어서,

상기 인증 데이터는 상기 최종 사용자 디바이스, 화자 정보 및 생체 인증 정보와 연관된 최종 사용자 디바이스 정보, 메타데이터 중 적어도 하나를 포함하는, 시스템.

청구항 68

제59항에 있어서,

상기 컴퓨터는,

상기 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하고;

상기 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 상기 하나 이상의 음향 특징들에 기초하여 스푸핑 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하도록 더 구성되는, 시스템.

청구항 69

제59항에 있어서,

상기 컴퓨터는 미디어 콘텐트 데이터베이스에서 상기 하나 이상의 화자 특성들에 대응하는 하나 이상의 콘텐트 특성들을 갖는 하나 이상의 미디어 콘텐트 파일들을 식별하도록 더 구성되는, 시스템.

청구항 70

제69항에 있어서,

상기 컴퓨터는 상기 화자의 상기 하나 이상의 화자 특성들 중 연령 특성이 연령 제한 콘텐트의 상기 하나 이상의 콘텐트 특성들의 해당 연령 제한 특성을 충족하는 것에 기초하여 상기 콘텐트 데이터베이스에서 상기 연령 제한 콘텐트를 활성화하도록 더 구성되는, 시스템.

청구항 71

방법에 있어서,

상기 컴퓨터에 의해, 화자에 대한 하나 이상의 임베딩들을 포함하는 상기 화자와 연관된 화자 프로파일을 획득하는 단계;

상기 컴퓨터에 의해, 상기 하나 이상의 임베딩들에 대한 하나 이상의 성숙 인자들 및 오수락률에 기초하여 상기 화자에 대한 성문에 대한 성숙 레벨을 결정하는 단계; 및

상기 컴퓨터에 의해, 상기 성숙 레벨 및 상기 하나 이상의 성숙 인자들에 따라 상기 화자 프로파일의 하나 이상의 유사도 임계값들을 업데이트하는 단계를 포함하는, 방법.

청구항 72

제71항에 있어서,

상기 화자 프로파일을 획득하는 단계는,

상기 컴퓨터에 의해, 상기 화자가 미디어 콘텐트 시스템의 새로운 사용자라고 결정하는 것에 응답하여 상기 화자 프로파일을 생성하는 단계를 포함하는, 방법.

청구항 73

제71항에 있어서,

상기 화자 프로파일을 획득하는 단계는,

상기 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 아키텍처를 적용함으로써 상기 화자에 대한 인바운드 임베딩을 추출하는 단계;

상기 컴퓨터에 의해, 상기 인바운드 임베딩과 상기 성문 사이의 상대적 거리에 기초하여 상기 인바운드 임베딩에 대한 유사도 점수를 생성하는 단계; 및

상기 컴퓨터에 의해, 상기 인바운드 임베딩에 대한 상기 유사도 점수가 상기 화자 프로파일의 상기 하나 이상의 유사도 임계값들 중 적어도 하나의 유사도 임계값을 충족한다고 결정하는 것에 응답하여 화자 데이터베이스에 저장된 상기 화자 프로파일을 식별하는 단계를 포함하는, 방법.

청구항 74

제73항에 있어서,

상기 적어도 하나의 임계값은 낮은 유사도 임계값을 포함하고, 상기 하나 이상의 임계값들은 상기 낮은 유사도 임계값 및 높은 유사도 임계값을 포함하는, 방법.

청구항 75

제73항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 임베딩에 대한 상기 유사도 점수가 상기 하나 이상의 유사도 임계값들 중 높은 유사도 임계값을 충족한다고 결정하는 것에 응답하여 상기 인바운드 임베딩에 따라 상기 화자에 대한 상기 성문을 업데이트하는 단계를 더 포함하며, 상기 컴퓨터는 상기 성문을 업데이트한 후 상기 성문에 대한 상기 성숙 레벨을 결정하는, 방법.

청구항 76

제75항에 있어서,

상기 컴퓨터에 의해, 상기 성문을 업데이트하는 데 사용되는 상기 인바운드 임베딩에 기초하여 상기 성문에 대한 상기 하나 이상의 성숙 인자들을 업데이트하는 단계를 더 포함하며, 상기 컴퓨터는 상기 하나 이상의 성숙 인자들을 업데이트한 후 상기 성문에 대한 상기 성숙 레벨을 결정하는, 방법.

청구항 77

제71항에 있어서,

상기 컴퓨터가 상기 성문의 상기 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

상기 컴퓨터에 의해, 상기 화자와 연관된 추가 인바운드 임베딩을 요청하는 프롬프트를 생성하는 단계; 및

각 추가 인바운드 임베딩이 상기 하나 이상의 유사도 임계값들을 충족하는 경우,

상기 컴퓨터에 의해, 상기 추가 인바운드 임베딩에 따라 상기 성문을 업데이트하는 단계; 및

상기 컴퓨터에 의해, 각 추가 인바운드 임베딩에 기초하여 상기 성문에 대한 상기 하나 이상의 성숙 인자들을 업데이트하는 단계를 더 포함하는, 방법.

청구항 78

제71항에 있어서,

상기 하나 이상의 유사도 임계값들을 업데이트하는 단계는,

상기 컴퓨터에 의해, 상기 성문에 대한 상기 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 상기 화자 프로파일의 상기 하나 이상의 임계값들을 증가시키는 단계를 포함하는, 방법.

청구항 79

제71항에 있어서,

상기 하나 이상의 성숙 인자들은 상기 화자의 임베딩 수, 상기 하나 이상의 임베딩들에서 발생하는 순 음성(net speech)의 지속시간, 및 상기 하나 이상의 임베딩들로부터의 오디오 품질 중 적어도 하나를 포함하는, 방법.

청구항 80

제71항에 있어서,

상기 컴퓨터에 의해, 관리 컴퓨터로부터 수신된 구성 입력에 기초하여 상기 오수락률을 결정하는 단계를 더 포함하는, 방법.

청구항 81

시스템에 있어서,

복수의 화자 프로파일들을 저장하도록 구성된 비일시적 기계 판독 가능 매체를 포함하는 화자 프로파일 데이터베이스; 및

컴퓨터로서,

화자에 대한 하나 이상의 임베딩들을 포함하는 상기 화자와 연관된 화자 프로파일을 획득하고;

상기 하나 이상의 임베딩들에 대한 하나 이상의 성숙 인자들 및 오수락률에 기초하여 상기 화자에 대한 성문에 대한 성숙 레벨을 결정하고;

상기 성숙 레벨 및 상기 하나 이상의 성숙 인자들에 따라 상기 화자 프로파일의 하나 이상의 유사도 임계값들을 업데이트하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함하는, 시스템.

청구항 82

제81항에 있어서,

상기 화자 프로파일을 획득하기 위해, 상기 컴퓨터는,

상기 화자가 미디어 콘텐트 시스템의 새로운 사용자라고 결정하는 것에 응답하여 상기 화자 프로파일을 생성하도록 구성되는, 시스템.

청구항 83

제81항에 있어서,

상기 화자 프로파일을 획득하기 위해, 상기 컴퓨터는,

인바운드 오디오 신호에 기계 학습 아키텍처를 적용함으로써 상기 화자에 대한 인바운드 임베딩을 추출하고;

상기 인바운드 임베딩과 상기 성문 사이의 상대적 거리에 기초하여 상기 인바운드 임베딩에 대한 유사도 점수를

생성하고;

상기 인바운드 임베딩에 대한 상기 유사도 점수가 상기 화자 프로파일의 상기 하나 이상의 유사도 임계값들 중 적어도 하나의 유사도 임계값을 충족한다고 결정하는 것에 응답하여 화자 데이터베이스에 저장된 상기 화자 프로파일을 식별하도록 더 구성되는, 시스템.

청구항 84

제83항에 있어서,

상기 적어도 하나의 임계값은 낮은 유사도 임계값을 포함하고, 상기 하나 이상의 임계값들은 상기 낮은 유사도 임계값 및 높은 유사도 임계값을 포함하는, 시스템.

청구항 85

제83항에 있어서,

상기 컴퓨터는,

상기 인바운드 임베딩에 대한 상기 유사도 점수가 상기 하나 이상의 유사도 임계값들 중 높은 유사도 임계값을 충족한다고 결정하는 것에 응답하여 상기 인바운드 임베딩에 따라 상기 화자에 대한 상기 성문을 업데이트하도록 더 구성되고, 상기 컴퓨터는 상기 성문을 업데이트한 후 상기 성문에 대한 상기 성숙 레벨을 결정하는, 시스템.

청구항 86

제85항에 있어서,

상기 컴퓨터는,

상기 성문을 업데이트하는 데 사용되는 상기 인바운드 임베딩에 기초하여 상기 성문에 대한 상기 하나 이상의 성숙 인자들을 업데이트하도록 더 구성되며, 상기 컴퓨터는 상기 하나 이상의 성숙 인자들을 업데이트한 후 상기 성문에 대한 상기 성숙 레벨을 결정하는, 시스템.

청구항 87

제81항에 있어서,

상기 컴퓨터는,

상기 컴퓨터가 상기 성문의 상기 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

상기 화자와 연관된 추가 인바운드 임베딩을 요청하는 프롬프트를 생성하고;

각 추가 인바운드 임베딩이 상기 하나 이상의 유사도 임계값들을 충족하는 경우,

상기 추가 인바운드 임베딩에 따라 상기 성문을 업데이트하고;

각 추가 인바운드 임베딩에 기초하여 상기 성문에 대한 상기 하나 이상의 성숙 인자들을 업데이트하도록 더 구성되는, 시스템.

청구항 88

제81항에 있어서,

상기 하나 이상의 유사도 임계값들을 업데이트하기 위해, 상기 디바이스는,

상기 성문에 대한 상기 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 상기 화자 프로파일의 상기 하나 이상의 임계값들을 증가시키도록 더 구성되는, 시스템.

청구항 89

제81항에 있어서,

상기 하나 이상의 성숙 인자들은 상기 화자의 임베딩 수, 상기 하나 이상의 임베딩들에서 발생하는 순 음성의

지속시간, 및 상기 하나 이상의 임베딩들로부터의 오디오 품질 중 적어도 하나를 포함하는, 시스템.

청구항 90

제81항에 있어서,

상기 컴퓨터는,

관리 컴퓨터로부터 수신된 구성 입력에 기초하여 상기 오수락률을 결정하도록 더 구성되는, 시스템.

청구항 91

디바이스 구현 방법에 있어서,

디바이스에 의해, 인바운드 화자의 발화를 포함하는 인바운드 오디오 신호를 수신하는 단계;

상기 디바이스에 의해, 상기 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 상기 인바운드 화자에 대한 인바운드 임베딩을 추출하는 단계;

상기 디바이스에 의해, 상기 인바운드 임베딩과 비일시적 기계 판독 가능 매체에 저장된 하나 이상의 성문들 사이의 상대적 거리에 따라 상기 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하는 단계;

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 성문을 사용하여 생성된 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여 상기 화자에 대한 상기 성문과 연관된 화자 식별자를 식별하는 단계; 및

상기 디바이스에 의해, 상기 화자 식별자를 콘텐트 서버에 전송하는 단계를 포함하는, 방법.

청구항 92

제91항에 있어서,

상기 디바이스는 스마트 TV, 셋톱 박스, 에지 디바이스, 리모콘, 및 이동 통신 디바이스 중 적어도 하나인, 방법.

청구항 93

제91항에 있어서,

상기 디바이스에 의해, 상기 디바이스에 결합된 디스플레이 디바이스를 통해 디스플레이하기 위해 상기 콘텐트 서버로부터 미디어 콘텐트를 수신하는 단계를 더 포함하는, 방법.

청구항 94

제91항에 있어서,

상기 디바이스는 상기 화자 식별자와 함께 미디어 콘텐트에 대한 요청을 상기 콘텐트 서버에 전송하는, 방법.

청구항 95

제91항에 있어서,

상기 디바이스에 의해, 상기 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하는 단계를 더 포함하며, 하나 이상의 유형의 생체 인식 특징들을 갖는 상기 특징들은 적어도 하나의 오디오 특징을 포함하는, 방법.

청구항 96

제91항에 있어서,

상기 디바이스의 마이크에 의해, 복수의 인바운드 화자들로부터 복수의 발화들을 포함하는 오디오 파형들을 수신하는 단계; 및

상기 디바이스에 의해, 상기 오디오 파형들을 상기 복수의 발화들을 포함하는 상기 인바운드 오디오 신호로 변환하는 단계를 더 포함하는, 방법.

청구항 97

제91항에 있어서,

상기 인바운드 오디오 신호는 복수의 인바운드 화자들의 복수의 발화들을 포함하고,

상기 디바이스는 상기 인바운드 오디오 신호에 상기 임베딩 추출 모델을 적용하여 상기 복수의 인바운드 화자들에 대응하는 상기 복수의 인바운드 임베딩들에 대한 복수의 유사도 점수들을 생성하고, 상기 복수의 유사도 점수들은 상기 디바이스의 비일시적 기계 판독 가능 스토리지에 저장된 복수의 성문들에 따라 생성되는, 방법.

청구항 98

제97항에 있어서,

상기 디바이스가 상기 복수의 임베딩들 중 제2 인바운드 임베딩에 대한 상기 복수의 유사도 점수들 중 제2 유사도 점수가 적어도 하나의 유사도 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

상기 디바이스에 의해, 상기 비일시적 기계 판독 가능 스토리지에 상기 제2 인바운드 임베딩 및 제2 화자 식별자를 포함하는 상기 제2 인바운드 화자에 대한 제2 화자 프로파일을 생성하는 단계를 더 포함하는, 방법.

청구항 99

제91항에 있어서,

상기 디바이스에 의해, 상기 인바운드 화자에 대한 하나 이상의 등록 오디오 신호들을 요청하는 상기 인바운드 화자에 디스플레이하기 위한 하나 이상의 등록 프롬프트들을 생성하는 단계;

상기 디바이스에 의해, 상기 하나 이상의 등록 오디오 신호들로부터 추출된 하나 이상의 등록 임베딩들에 기초하여 상기 인바운드 화자에 대한 상기 성문을 생성하는 단계;

상기 디바이스가 상기 성문의 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

상기 디바이스에 의해, 제2 등록 오디오 신호를 요청하는 제2 등록 프롬프트를 생성하는 단계; 및

상기 디바이스에 의해, 상기 제2 등록 오디오 신호로부터 추출된 제2 등록 임베딩에 기초하여 상기 인바운드 화자에 대한 상기 성문을 업데이트하는 단계를 더 포함하는, 방법.

청구항 100

제91항에 있어서,

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호로부터 하나 이상의 음향 특징들을 추출하는 단계; 및

상기 컴퓨터에 의해, 상기 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 상기 하나 이상의 음향 특징들에 기초하여 스푸핑 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하는 단계를 더 포함하는, 방법.

청구항 101

시스템에 있어서,

화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및

디바이스로서,

인바운드 화자의 발화를 포함하는 인바운드 오디오 신호를 수신하고;

상기 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 상기 인바운드 화자에 대한 인바운드 임베딩을 추출하고;

상기 인바운드 임베딩과 상기 화자 데이터베이스에 저장된 하나 이상의 성문들 사이의 상대적 거리에 따라 상기 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하고;

상기 컴퓨터에 의해, 상기 인바운드 화자에 대한 성문을 사용하여 생성된 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여 상기 화자에 대한 상기 성문과 연관된 화자 식별자를 식별하고;

상기 화자 식별자를 콘텐트 서버에 전송하도록 구성된 프로세서를 포함하는, 상기 디바이스를 포함하는, 시스템.

청구항 102

제101항에 있어서,

상기 디바이스는 스마트 TV, 셋톱 박스, 에지 디바이스, 리모콘, 및 이동 통신 디바이스 중 적어도 하나인, 시스템.

청구항 103

제101항에 있어서,

상기 디바이스는 상기 디바이스에 결합된 디스플레이 디바이스를 통해 디스플레이하기 위해 상기 콘텐트 서버로부터 미디어 콘텐트를 수신하도록 더 구성되는, 시스템.

청구항 104

제101항에 있어서,

상기 디바이스는 상기 화자 식별자와 함께 미디어 콘텐트에 대한 요청을 상기 콘텐트 서버에 전송하도록 구성되는, 시스템.

청구항 105

제101항에 있어서,

상기 디바이스는,

상기 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하도록 더 구성되며, 하나 이상의 유형의 생체 인식 특징들을 갖는 상기 특징들은 적어도 하나의 오디오 특징을 포함하는, 시스템.

청구항 106

제101항에 있어서,

복수의 인바운드 화자들로부터 복수의 발화들을 포함하는 오디오 파형들을 수신하도록 구성된 마이크를 더 포함하며;

상기 디바이스는 상기 오디오 파형들을 상기 복수의 발화들을 포함하는 상기 인바운드 오디오 신호로 변환하도록 더 구성되는, 시스템.

청구항 107

제101항에 있어서,

상기 인바운드 오디오 신호는 복수의 인바운드 화자들의 복수의 발화들을 포함하며,

상기 디바이스는 상기 인바운드 오디오 신호에 상기 임베딩 추출 모델을 적용하여 상기 복수의 인바운드 화자들에 대응하는 상기 복수의 인바운드 임베딩들에 대한 복수의 유사도 점수들을 생성하도록 구성되며, 상기 디바이스는 상기 화자 데이터베이스에 저장된 복수의 성문들에 따라 상기 복수의 유사도 점수들을 생성하는, 시스템.

청구항 108

제107항에 있어서,

상기 디바이스는,

상기 복수의 임베딩들 중 제2 인바운드 임베딩에 대한 상기 복수의 유사도 점수들 중 제2 유사도 점수가 적어도

하나의 유사도 임계값을 충족하지 못한다고 결정하고;

상기 비일시적 기계 판독 가능 스토리지에 상기 제2 인바운드 임베딩 및 제2 화자 식별자를 포함하는 상기 제2 인바운드 화자에 대한 제2 화자 프로파일을 생성하도록 더 구성되는, 시스템.

청구항 109

제101항에 있어서,

상기 디바이스는,

상기 인바운드 화자에 대한 하나 이상의 등록 오디오 신호들을 요청하는 상기 인바운드 화자에 디스플레이하기 위한 하나 이상의 등록 프롬프트들을 생성하고;

상기 하나 이상의 등록 오디오 신호들로부터 추출된 하나 이상의 등록 임베딩들에 기초하여 상기 인바운드 화자에 대한 상기 성문을 생성하고;

상기 디바이스가 상기 성문의 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여,

제2 등록 오디오 신호를 요청하는 제2 등록 프롬프트를 생성하고;

상기 제2 등록 오디오 신호로부터 추출된 제2 등록 임베딩에 기초하여 상기 인바운드 화자에 대한 상기 성문을 업데이트하도록 더 구성되는, 시스템.

청구항 110

제101항에 있어서,

상기 디바이스는,

상기 인바운드 오디오 신호로부터 하나 이상의 음향 특징들을 추출하고;

상기 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 상기 하나 이상의 음향 특징들에 기초하여 스푸핑 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하도록 더 구성되는, 시스템.

발명의 설명

기 술 분 야

[0001] **관련 출원들에 대한 상호 참조**

[0002] 본 출원은 2020년 4월 15일에 출원된 미국 임시 출원 번호 제63/010,504호에 대한 우선권을 주장하며, 이는 그 전체가 참조로 통합된다.

[0003] 본 출원은 일반적으로 2016년 9월 12일자로 출원된 "End-To-End Speaker Recognition Using Deep Neural Network"이라는 명칭의 미국 출원 번호 제15/262,748호에 관한 것으로, 미국 특허 번호 제9,824,692호로 발행되며, 이는 그 전체가 본원에 참조로 통합된다.

[0004] 본 출원은 일반적으로 2018년 2월 7일자로 출원된 "Age Compensation in Biometric Systems Using Time-Interval, Gender and Age"이라는 명칭의 미국 출원 번호 제15/262,748호에 관한 것으로, 미국 특허 번호 제10,672,403호로 발행되며, 이는 그 전체가 본원에 참조로 통합된다.

[0005] 본 출원은 일반적으로 2018년 3월 2일자로 출원된 "Method and Apparatus for Detecting Spoofing Conditions"이라는 명칭의 미국 출원 번호 제15/910,387호에 관한 것으로, 미국 특허 번호 제10,692,502호로 발행되며, 이는 그 전체가 본원에 참조로 통합된다.

[0006] 본 출원은 일반적으로 2021년 1월 22일자로 출원된 "Robust Spoofing Detection System Using Deep Residual Neural Networks"이라는 명칭의 미국 출원 번호 제17/155,851호에 관한 것으로, 이는 그 전체가 본원에 참조로 통합된다.

[0007] 본 출원은 일반적으로 2021년 3월 4일자로 출원된 "Systems and Methods of Speaker-Independent Embedding

for Identification and Verification from Audio"이라는 명칭의 미국 출원 번호 제17/192,464호에 관한 것으로, 이는 그 전체가 본원에 참조로 통합된다.

[0008] **기술분야**

[0009] 본 출원은 일반적으로 오디오 처리 신경망을 훈련시키고 배포하기 위한 시스템들 및 방법들에 관한 것이다.

배 경 기 술

[0010] 사물 인터넷(Internet of Things; IoT) 디바이스들의 출현은 음성 명령들과 상호 작용하는 기계들의 새로운 채널들로 이어졌다. 종종 디바이스들과의 많은 상호 작용들에는 개인 데이터와 민감한 데이터에 대한 동작들을 수행하는 것이 수반된다. 많은 새로운 모바일 앱들과 가정용 개인 어시스턴트들이 디바이스와의 음성 기반 상호작용들을 사용하여 금융 거래를 가능하게 하고 있다. 콜 센터, 특히 콜 센터의 상담원과의 상호작용들은 더 이상 중요한 개인 정보를 관리하는 기관들의 음성 기반 상호작용들의 유일한 사례가 아니다. 다양한 에지(edge) 디바이스들이나 IoT 디바이스들을 운용하거나, 콜 센터들에 연락함으로써 사용자 계정들에 접근 및 관리하는 발신자/화자의 ID를 일정한 레벨의 정확도와 보안 수준에 따라 신뢰성있게 검증하는 것이 필수적이다.

[0011] 자동 음성 인식(Automatic Speech Recognition; ASR) 및 자동 화자 확인(Automatic Speaker Verification; ASV) 시스템들은 보안 및 인증 기능들뿐만 아니라 기타 음성 기반 동작들에 종종 사용된다. 음성 생체 인식의 대부분의 구현들은 능동적 및 정적 등록을 사용하며 일반적으로 오디오 발화(utterance)들과 화자 ID 사이의 링크가 알려져 있다고 가정한다. 능동적 등록은 일반적으로 음성 생체 인식 시스템에 의해 정의된 기준이 충족될 때까지 패스프레이즈(passphrase)를 반복하거나 자유롭게 말해야 하는 등록 단계에 대해 사용자에게 프롬프트가 표시되는 경우이다. 능동적 등록은 사용자가 처음에 각자의 디바이스를 설정하거나 오버-더-톱(over-the-top) 서비스를 사용하기 시작할 때 정적 등록과 함께 종종 발생한다. 능동적 등록은 시간이 많이 소요될 수 있으며, 사용자들이 등록을 하지 않을 수 있기 때문에 음성 생체 인식 배포가 성공적이지 않을 수 있다. 또한, 정적 등록은 더 많은 사람들이 음성 생체 인식 시스템을 사용하기를 원하거나 음성이 변경됨에 따라 음성 모델들이 오래되거나 부정확한 검증을 초래할 수 있다.

[0012] 게다가, 오버-더-톱(over-top; OTT) 서비스들은 은행과 같은 자동 음성 검증을 사용하는 다른 서비스들과 다를 수 있는데, OTT 서비스는 화자가 화자의 프로파일과 연관된 미리 정의된 기준을 충족하는지 여부를 단독으로 결정(예를 들어, 화자의 음성이 임의의 다른 화자 프로파일들에 관계없이 화자의 프로파일에 해당하는 음성과 일치하는지 여부를 결정)하는 것과는 대조적으로 한 번에 다수의 화자들로부터 개별 화자들의 식별들을 요구할 수 있기 때문이다. 특히 다수의 화자들이 동시에 또는 간헐적으로 서로 말하는 경우, 화자들 사이를 능동적으로 구별할 수 있는 시스템을 유지하기 어려울 수 있다. 또한, 다수의 사람들이 함께 모여 비디오들을 보거나 함께 음악을 듣거나 함께 차에 탈 때와 같이, 시스템이 한 번에 다수의 개인들로부터의 음성을 식별할 때 콘텐트를 제공하거나 개인용 에지 디바이스를 구성하는 것이 어려울 수 있다.

[0013] 따라서 필요한 것은 서비스가 동작할 때 새로운 화자들을 등록하고 시스템이 발화 일치를 위해 프로파일들을 서로 구별할 수 있도록 사전 설정 및/또는 설정되지 않은 화자 프로파일들을 사용하여 콘텐트를 제공하거나 화자용 에지 디바이스를 구성하는 개선된 접근 방식들이다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0014] 전술한 단점들을 해결할 수 있고 또한 임의 개수의 추가적 또는 대안적 이익들 및 이점들을 제공할 수 있는 시스템들 및 방법들이 본원에 개시된다. 본원에 설명된 실시예들은 수동적, 능동적, 지속적 또는 정적 동작들, 또는 이들의 일부 하이브리드 조합이 가능한 유연한 음성 생체 인식 시스템을 제공한다. 특히, 본원에 설명된 시스템들 및 방법들은 수동적으로 및/또는 지속적으로, 일부 경우에는 능동적으로 및/또는 정적으로 추가하여, 화자들이 에지 디바이스(예를 들어, 자동차, 텔레비전, 라디오, 전화) 내부 또는 주변에서 말할 때 화자들을 등록한다. 이러한 시스템들 및 방법들을 구현함으로써, 디바이스는 새로운 화자가 프로파일을 설정하기를 원할 때마다 디바이스를 능동적으로 재구성하기 위해 새로운 화자가 프롬프트된 발화를 미러링할 필요 없이 즉석에서

새로운 사용자들을 식별할 수 있다. 시스템들 및 방법들은 화자들이 시스템에 발화들을 제공하여 화자들에 대한 최신 프로파일들을 유지하고 잘못된 인증 수락들 및/또는 거부들을 방지할 때 화자 프로파일들을 구성 및 재구성하는 방법을 추가로 제공한다. 시스템들 및 방법들은 일부 경우에는 화자 ID들에 대한 지식이 전혀 없을 가능성이 있는 수동적 및 지속적 음성 생체 인식 시스템을 제공한다. 시스템들 및 방법들은 비지도 방식으로 ID들을 생성할 수 있으며, 일부 경우에는 시스템이 임의의 저장된 사용자 프로파일들에 대한 기준을 충족하지 않는 화자들을 식별할 때 개별 화자들을 수동적으로 등록하고 인식할 수 있다. 이러한 시스템들 및 방법들은 오버-더-톱 서비스들(예를 들어, 관련 콘텐트를 제공하기 위해 텔레비전 시청자 식별) 및/또는 콜 센터 사용 사례에 대한 IoT들(예를 들어, 자동차의 운전자를 식별 및 운전자의 화자 프로파일과 연관된 설정들에 기초하여 자동차의 설정들을 구성)을 포함하는, 광범위한 애플리케이션들에 걸쳐 개인화 및 보안 목적으로 사용될 수 있다.

[0015] 일 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩(inbound embedding)을 추출하는 단계; 컴퓨터에 의해, 화자 프로파일 데이터베이스의 화자 프로파일에 저장된 성문(voiceprint)과 인바운드 임베딩 사이의 거리에 기초하여 유사도 점수(similarity score)를 생성하는 단계; 및 컴퓨터가 인바운드 임베딩에 대한 유사도 점수가 유사도 임계값에 충족하지 못한다고 결정하는 것에 응답하여, 컴퓨터에 의해, 화자 프로파일 데이터베이스에, 인바운드 임베딩을 포함하는 인바운드 화자에 대한 새로운 화자 프로파일을 생성하는 단계로서, 새로운 화자 프로파일은 인바운드 임베딩을 새로운 성문으로 저장하는 데이터베이스 레코드인, 상기 생성하는 단계를 포함한다.

[0016] 다른 실시예에서, 시스템은 화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 컴퓨터로서, 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하고; 화자 프로파일 데이터베이스의 화자 프로파일에 저장된 성문과 인바운드 임베딩 사이의 거리에 기초하여 유사도 점수를 생성하고; 컴퓨터가 인바운드 임베딩에 대한 유사도 점수가 유사도 임계값을 충족하지 못한다고 결정하는 것에 응답하여, 화자 프로파일 데이터베이스에, 인바운드 임베딩을 포함하는 인바운드 화자에 대한 새로운 화자 프로파일을 생성하되, 새로운 화자 프로파일은 상기 인바운드 임베딩을 새로운 성문으로 저장하는 데이터베이스 레코드인, 상기 생성하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함한다.

[0017] 다른 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하는 단계; 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 아키텍처를 적용하여 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들을 추출하는 단계; 복수의 인바운드 화자들의 각 인바운드 화자에 대해, 컴퓨터에 의해, 상기 인바운드 화자에 대한 인바운드 임베딩에 기초하여 하나 이상의 유사도 점수들을 생성하는 단계로서, 인바운드 화자에 대한 각 유사도 점수는 인바운드 임베딩과 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들 사이의 거리를 나타내는, 상기 생성하는 단계; 및 컴퓨터에 의해, 하나 이상의 성문들로부터 인바운드 화자에 대한 가장 가까운 성문을 식별하는 단계로서, 가장 가까운 성문은 인바운드 화자에 대해 생성된 하나 이상의 유사도 점수들 중 최대 유사도 점수에 대응하는, 상기 식별하는 단계; 및 각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족하는 경우, 컴퓨터에 의해, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하는 단계를 포함한다.

[0018] 다른 실시예에서, 시스템은 화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 컴퓨터로서, 복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하고; 인바운드 오디오 신호에 기계 학습 아키텍처를 적용하여 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들을 추출하고; 복수의 인바운드 화자들의 각 인바운드 화자에 대해, 인바운드 화자에 대한 인바운드 임베딩에 기초하여 하나 이상의 유사도 점수들을 생성하되, 인바운드 화자에 대한 각 유사도 점수는 인바운드 임베딩과 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들 사이의 거리를 나타내고; 하나 이상의 성문들로부터 인바운드 화자에 대한 가장 가까운 성문을 식별하되, 가장 가까운 성문은 인바운드 화자에 대해 생성된 하나 이상의 유사도 점수들 중 최대 유사도 점수에 대응하고; 각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족하는 경우, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함한다.

[0019] 다른 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 콘텐트 서버를 통해 최종 사용자 디바이스로부터 인바운드 화자에 대한 인바운드 오디오 신호를 수신하는 단계; 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하는 단계; 컴퓨터에 의해, 인바운드 임베딩과 화

자 데이터베이스의 화자 프로파일에 저장된 성문 사이의 거리에 기초하여 인바운드 임베딩에 대한 유사도 점수를 생성하는 단계로서, 유사도 점수는 하나 이상의 유사도 점수 임계값들을 충족하는, 상기 생성하는 단계; 컴퓨터에 의해, 화자 프로파일에서 상기 콘텐트 서버에 대한 하나 이상의 콘텐트 특성들에 대응하는 하나 이상의 화자 특성들을 식별하는 단계; 및 컴퓨터에 의해, 인바운드 화자와 연관된 하나 이상의 화자 특성들을 미디어 콘텐트 서버에 전송하는 단계를 포함한다.

[0020] 다른 실시예에서, 시스템은 복수의 화자 프로파일들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 서버로서, 콘텐트 서버를 통해 최종 사용자 디바이스로부터 인바운드 화자에 대한 인바운드 오디오 신호를 수신하고; 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하고; 인바운드 임베딩과 상기 화자 데이터베이스의 화자 프로파일에 저장된 성문 사이의 거리에 기초하여 인바운드 임베딩에 대한 유사도 점수를 생성하되, 유사도 점수는 하나 이상의 유사도 점수 임계값들을 충족하고; 화자 프로파일에서 상기 콘텐트 서버에 대한 하나 이상의 콘텐트 특성들에 대응하는 하나 이상의 화자 특성들을 식별하고; 인바운드 화자와 연관된 하나 이상의 화자 특성들을 미디어 콘텐트 서버에 전송하도록 구성된 프로세서를 포함하는, 상기 서버를 포함한다.

[0021] 다른 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 화자에 대한 하나 이상의 임베딩들을 포함하는 화자와 연관된 화자 프로파일을 획득하는 단계; 컴퓨터에 의해, 하나 이상의 임베딩들에 대한 하나 이상의 성숙(maturity) 인자들 및 오수락률에 기초하여 화자에 대한 성문에 대한 성숙 레벨을 결정하는 단계; 컴퓨터에 의해, 성숙 레벨 및 하나 이상의 성숙 인자들에 따라 화자 프로파일의 하나 이상의 유사도 임계값들을 업데이트하는 단계를 포함한다.

[0022] 다른 실시예에서, 시스템은 복수의 화자 프로파일들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 프로파일 데이터베이스; 및 컴퓨터로서, 화자에 대한 하나 이상의 임베딩들을 포함하는 화자와 연관된 화자 프로파일을 획득하고; 하나 이상의 임베딩들에 대한 하나 이상의 성숙 인자들 및 오수락률에 기초하여 화자에 대한 성문에 대한 성숙 레벨을 결정하고; 성숙 레벨 및 하나 이상의 성숙 인자들에 따라 화자 프로파일의 하나 이상의 유사도 임계값들을 업데이트하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함한다.

[0023] 다른 실시예에서, 디바이스 구현 방법은, 디바이스에 의해, 인바운드 화자의 발화를 포함하는 인바운드 오디오 신호를 수신하는 단계; 디바이스에 의해, 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하는 단계; 디바이스에 의해, 인바운드 임베딩과 비일시적 기계 판독 가능 매체에 저장된 하나 이상의 성문들 사이의 상대적 거리에 따라 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하는 단계; 컴퓨터에 의해, 성문을 사용하여 생성된 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여 인바운드 화자에 대한 성문과 연관된 화자 식별자를 식별하는 단계; 및 디바이스에 의해, 화자 식별자를 콘텐트 서버에 전송하는 단계를 포함한다.

[0024] 다른 실시예에서, 시스템은 화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 디바이스로서, 인바운드 화자의 발화를 포함하는 인바운드 오디오 신호를 수신하고; 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하고; 인바운드 임베딩과 화자 데이터베이스에 저장된 하나 이상의 성문들 사이의 상대적 거리에 따라 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하고; 컴퓨터에 의해, 성문을 사용하여 생성된 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여 인바운드 화자에 대한 성문과 연관된 화자 식별자를 식별하고; 화자 식별자를 콘텐트 서버에 전송하도록 구성된 프로세서를 포함하는, 상기 디바이스를 포함한다.

[0025] 전술한 일반적인 설명 및 다음의 상세한 설명은 모두 예시적이고 설명적이며 청구된 바와 같은 본 발명의 추가 설명을 제공하기 위한 것으로 이해되어야 한다.

도면의 간단한 설명

[0026] 본 개시는 하기의 도면들을 참조함으로써 더 잘 이해될 수 있다. 도면들의 컴포넌트들은 반드시 일정한 비율을 유지해야 하는 것은 아니며, 대신에 본 개시의 원리들을 설명하는 데 중점을 둔다. 이 도면들에서, 참조 부호들은 여러 도면들에 걸쳐 대응되는 부분들을 나타낸다.

도 1은 오디오 처리 기계 학습 동작들을 사용하는 시스템의 컴포넌트들을 도시한다.

도 2는 기계 학습 모델들 및 기타 기계 학습 아키텍처가 로컬 디바이스 상에 구현되는, 오디오 처리 기계 학습

동작들을 사용하는 시스템의 컴포넌트들을 도시한다.

도 3a 내지 3b는 능동적으로 사용자들을 등록하고 사용자들을 식별(인증)하기 위한 방법의 동작 단계들을 도시한다.

도 4는 오디오 처리 시스템에서 적응적 임계화를 위한 방법의 동작 단계들을 도시한다.

도 5는 오디오 처리 시 강한 발화와 약한 발화를 식별하고 평가하기 위한 방법의 실행 단계들을 도시한다.

도 6은 오디오 처리 시 화자들을 클러스터링하기 위한 방법의 동작 단계들을 도시한다.

도 7a는 현재 및/또는 이력 정보에 따라 하나 이상의 성문들의 라벨 식별자들(예를 들어, 화자 식별자들, 가입자 식별자들)을 정정하기 위한 방법의 동작 단계들을 도시한다.

도 7b는 특정 화자에 대한 클러스터들 및 기타 추정 화자들의 클러스터들을 사용하는 라벨 정정의 일 예를 예시한다.

도 8은 수동적 및 지속적 등록 구성들을 사용하여 오디오 처리를 위한 방법의 동작 단계들을 도시한다.

도 9는 혼합된 능동적-수동적 및 지속적 등록 구성들을 사용하여 오디오 신호들을 오디오 처리하기 위한 방법의 동작 단계들을 도시한다.

도 10은 능동적 및 지속적 등록 구성들을 사용하여 오디오 신호들을 오디오 처리하기 위한 방법의 동작 단계들을 도시한다.

도 11a 내지 11b는 기계 학습 모델들이 차량에 의해 구현되는, 오디오 처리 기계 학습 동작들을 사용하는 시스템의 컴포넌트들을 도시한다.

도 12는 성숙 인자 및 오수락률에 기초한 유사도 임계값 스케줄러에 대한 예시적인 테이블을 예시한다.

발명을 실시하기 위한 구체적인 내용

[0027] 이제 도면들에 예시된 예시적인 실시예들에 대한 참조가 이루어질 것이며, 특정 언어는 동일한 것을 설명하기 위해 본원에서 사용될 것이다. 그럼에도 불구하고 이에 의해 본 발명의 범위가 제한되지 않음이 이해될 것이다. 본원에 예시된 본 발명의 특징들에 대한 변경들 및 추가 수정들, 그리고 관련 기술 분야의 숙련자 및 본 개시를 소유한 자에게 발생할 수 있는 본원에 예시된 바와 같은 본 발명의 원리들의 추가 적용은 본 발명의 범위 내에서 고려되어야 한다.

[0028] 화자 인식 및 기타 동작들(예를 들어, 인증)을 위한 음성 생체 인식은 일반적으로 화자 샘플들 및 특정 화자의 샘플들의 세계에서 생성된 모델들 또는 특징 벡터들('임베딩(embedding)들'이라고도 함)에 의존한다. 일 예로서, 훈련 단계(또는 재훈련 단계) 동안, 서버 또는 기타 컴퓨팅 디바이스는 복수의 훈련 오디오 신호들을 사용하여 음성 인스턴스들을 인식하고 구별하도록 훈련된 음성 인식 엔진(예를 들어, 인공 지능 및/또는 기계 학습 프로그래밍 소프트웨어)을 실행한다. 기계 학습 아키텍처는 해당 입력들에 따라 특정 결과들을 출력하고 예상된 출력과 관찰된 출력을 비교하여 손실 함수에 따라 결과들을 평가한다. 그런 다음 훈련 동작들은 (기계 학습 아키텍처의 신경망의) 가중 값들 또는 하이퍼-파라미터들을 조정하고 예상된 출력들과 관찰된 출력들이 수렴할 때까지 입력들에 기계 학습 아키텍처를 다시 적용한다. 그런 다음 서버는 하이퍼-파라미터들을 수정하고, 일부 경우에는, 훈련에 사용되는 신경망 아키텍처의 하나 이상의 계층들을 비활성화한다.

[0029] 기계 학습 아키텍처를 훈련한 후, 서버는 특정 화자에 대한 등록 동작들 동안 특정 화자를 인식하기 위해 기계 학습 아키텍처를 추가로 개선하고 개발할 수 있다. 음성 인식 엔진은 화자의 발화(utterance)들을 갖는 등록자 오디오 신호들로부터 추출된 임베딩들을 사용하여 등록자에게 모델 임베딩(때로는 "성문(voiceprint)"이라고 함)을 생성할 수 있다. 추후 인바운드 오디오 신호들 동안, 서버는 추후 인바운드 오디오 신호로부터 추출된 인바운드 임베딩(inbound embedding)을 등록자의 성문과 일치시키는 것에 기초하여 추후 오디오 신호들이 알려진 화자를 포함하는지 여부를 확인하기 위해 화자 프로파일에 저장된 성문들을 참조한다.

[0030] 이러한 접근 방식들은 일반적으로 성공적이며 콜 센터에 대한 인바운드 전화 통화를 평가하는 맥락에서 등록자를 검출하는 데 적합하다. 사용자들이 예컨대 텔레비전을 시청하거나 특정 IoT 또는 음성 지원 디바이스들(예를 들어, 차량들, 스마트 기기들, 개인 어시스턴스들)을 동작시키는 경우와 같이, 보다 유동적이거나 덜 구조화된 경험을 선호할 때, 다른 맥락에서는 등록 및 배포 동작들에 대한 보다 유연하고 덜 가시적인 접근 방식들이 바

람직할 수 있다.

[0031] **성문들 및 콘텐트 서비스들을 위한 기계 학습 모델**

[0032] 본원에 설명된 실시예들은 음성 기반 인터페이스, 관리, 인증 및 콘텐트 개인화를 위한 음성 인식을 포함하는, 생체 인식을 위한 시스템들 및 방법들을 개시한다. 컴퓨팅 디바이스들은, 오디오 신호들을 처리하기 위해, 가우시안 매트릭스 모델들(Gaussian Matrix Models; GMM들) 및 신경망들을 포함하는, 다양한 유형의 기계 학습 아키텍처 계층들 또는 동작들을 구현하는 소프트웨어 프로그래밍을 실행한다. 기계 학습 아키텍처는 일반적으로 예를 들어 특징 벡터들을 생성하고 입력 오디오 신호의 양태들을 나타내거나 모델링하는 임베딩들을 추출하고; 임베딩들에 따른 분류를 수행하고; 특정 화자의 발화들로부터 추출된 하나 이상의 임베딩들에 기초하여 특정 화자에 대한 모델 임베딩(예를 들어, 성문)을 생성하고; 저장된/예상된 특징들 또는 과 비교하여 추출된 특징들 또는 임베딩들 사이의 유사도들 또는 차이들에 기초하여 화자들의 유사하게 들리는 발화들 또는 오디오 신호 특징들을 클러스터링하는 임의 개수의 기계 학습 모델들을 포함한다. 성문이 생성되면, 시스템은 발화들로부터 추출한 임베딩들을 성문과 비교하여 해당 발화가 (예를 들어, 성문으로 나타나는 화자가 말한) 성문과 연관되어야 할 가능성을 나타내는 유사도 점수(similarity score)를 생성하여 개인들의 발화들을 검출할 수 있다. 유사도 점수가 하나 이상의 미리 구성된 유사도 점수들을 충족하는 경우, 시스템은 발화에 대한 임베딩들을 화자와 연관시키고/시키거나 성문을 업데이트하기 위해 성문에 대한 발화에 대한 임베딩을 포함할 수 있다. 특정 실시예들의 설명은 훈련 동작들을 언급하고 있지만, 본원에 개시된 실시예들은 일반적으로 훈련 단계가 완료되고 등록 단계로 시작하는 것으로 가정한다.

[0033] **수동적 등록**

[0034] 본원에 설명된 음성 생체 인식 시스템들은 일반적으로 서버 또는 다른 컴퓨팅 디바이스가 훈련된 기계 학습 아키텍처를 포함한다고 가정한다. 이와 같이, 실시예들은 등록 및 배포 단계들의 음성 처리로 진행하며, 여기서 서버는 능동적 또는 수동적 등록 구성에서 등록 동작들을 수행한다. 능동적 등록 구성들에서, 오디오 또는 시각적 사용자 인터페이스(예를 들어, 전화, 텔레비전 화면)는 사용자에게 다양한 문구들을 말하도록 지시하는 프롬프트들을 제시한다. 프롬프트에 응답하여, 사용자는 예를 들어 하나 이상의 기준(예를 들어, 발화들의 수, 입력들 사이의 시간)이 충족될 때까지 패스프레이즈를 반복하거나 자유롭게 말할 수 있다. 관리 사용자 입력이 기준을 정의하거나, 하나 이상의 기계 학습 모델들이 기준을 자동으로 설정하거나 조정한다.

[0035] 수동적 등록에는 사용자 프롬프트들이 필요하지 않지만, 일부 구현들에서는 하이브리드 접근 방식들(예를 들어, 능동적 및 수동적)을 사용한다. 오히려, 서버는 특징을 추출하고, 분류를 수행하고, 기타 동작들을 수행하기 위해, 사용자 인식을 요구하지 않고, 오디오 신호들에 다양한 기계 학습 계층들을 수동적으로 적용한다. 유리하게는, 본원에 설명된 화자 모델링 동작들(예를 들어, 생체 인식 인증, 콘텐트 개인화, 화자 분할)을 사용하는 다운스트림 동작들이 끊김 없고 마찰이 없는 방식으로 발생하므로, 사용자가 상호 작용을 변경하거나 중단할 필요가 없어, 등록 또는 기타 동작들에 소요되는 시간을 절약할 수 있다.

[0036] **지속적 등록**

[0037] 서버는 등록 및 배포 단계 동안 정적 또는 지속적 음성 처리 동작들을 추가로 구현할 수 있다. 정적 동작들에서, 음성 생체 인식 시스템은 화자들을 한 번 또는 고정된 시간에 등록하고, 서버는 초기 등록 후 새로운 인입(incoming) 오디오 신호들로부터의 등록 동작들을 실행하거나 기계 학습 아키텍처 또는 모델들(예를 들어, 파라미터들, 성문들)을 업데이트하지 않는다.

[0038] 지속적 음성 처리 동작들을 구현함으로써, 음성 생체 인식 시스템은 새로운 인입 오디오 신호들을 사용하고 이로부터 이점들을 얻는다. 기계 학습 아키텍처는 초기에 정적 음성 처리 동작들을 구현하여 고정된 시간에 화자들을 능동적으로 등록할 수 있으며, 서버는 새로운 발화들을 추가로 수집하고 고정된 시간 이후의 시간에 새 화자들을 등록하고 검출하는 기계 학습 모델들을 개발할 수 있다. 지속적 음성 처리 동작들을 구현함으로써, 지속적 음성 처리 동작들은 새로운 발화들을 검출하고 새로운 발화들에 대해 추출된 임베딩들을 미리 정의된 기준(예를 들어, 성문들, 유사도 점수들, 인증 데이터, 사용자 정보, 디바이스 정보)과 비교하여 시간이 지남에 따라 새로운 화자들을 식별 및 등록할 수 있다. 기준은 예를 들어 생체 인식 정보, 화자/사용자 정보, 디바이스 정보 및 최종 사용자 디바이스들로부터의 데이터 입력들과 함께 수신된 메타데이터로부터 추출된 다양한 유형의 특징들을 포함할 수 있다. 지속적 등록 동작들을 수행하는 음성 생체 인식 시스템들은 다양한 특징들 및 기타 유형의 정보가 포함된 등록 입력들(예를 들어, 등록 오디오 신호들)을 수동적으로 캡처하고 분석할 수 있다. 그런 다음 시스템은 입력 오디오 신호의 특징들을 기반으로 새 화자들을 자동으로 검출하고 시스템에서 참조하는

새 성문(예를 들어, 모델 임베딩)을 생성하여 등록된 화자들을 식별하거나 알 수 없는 새 화자들을 구별한다. 이러한 시스템들은 부실을 피하기 위해 화자 프로파일들 및 화자 성문들의 지속적 등록 및 이에 대한 주기적 업데이트들을 추가로 수행할 수 있으며, 이로 인해 잘못된 거부 또는 수락 비율을 증가시킬 수 있다.

[0039] **조건 종속적 적응 임계화**

[0040] 화자 인식 시스템들의 일부 실시예들은 고정 임계화를 사용한다. 고정 임계화를 사용하는 단일-화자 검증의 경우, 서버는 인바운드 오디오 신호로부터 추출된 화자 임베딩을 등록된 임베딩(성문)과 비교하여 유사도 또는 예측 점수를 계산한다. 유사도 점수가 미리 정의된 화자-독립적 고정 임계값을 충족하면, 컴퓨팅 디바이스는 인바운드 오디오 샘플을 검증한다. 그렇지 않으면, 컴퓨팅 디바이스는 인바운드 오디오 신호를 거부하거나 잘못된 예측 점수를 보고한다. 화자의 명수(N)를 갖는 고정 임계값을 사용하는 다중-화자 검증 또는 오픈-세트(open-set) 식별의 경우, 컴퓨팅 디바이스는 인바운드 오디오 신호로부터 N개의 인바운드 화자 임베딩들을 추출하고 N개의 인바운드 임베딩들을 성문들(V)의 수와 비교하여 NxV 유사도 점수를 계산함으로써, 컴퓨팅 디바이스는 각 인바운드 임베딩을 각 성문과 비교한다. 그런 다음 컴퓨팅 디바이스는 N개의 상이한 유사도 점수들을 출력한다. 컴퓨팅 디바이스는 각 화자에 대한 최대 유사도 점수만을 고려하며, 여기서 최대 유사도 점수는 특정 화자 임베딩과 특정 성문 사이의 가장 근접한 일치를 나타낸다. 특정 화자 임베딩의 최대 유사도 점수가 미리 정의된 화자-독립적 고정 임계값을 충족하면, 컴퓨팅 디바이스는 다중-화자 오디오 샘플에서 해당 화자를 검증 또는 식별한다.

[0041] 일부 경우에, 고정 임계값 동작들을 사용하면 일부 성문들은 다른 모델들보다 더 빨리 성숙할 수 있다. 여러 성숙(maturation) 인자들에서 품질이 좋지 않은 메트릭들에 기초한 성문은 컴퓨팅 디바이스가 수락할 수 없는 비율로 화자들을 잘못 수락하도록 한다. 성숙 인자는 예를 들어, 등록 발화들의 수를 포함한다. 예를 들어, 50개의 발화들로 등록된 화자 성문은 단 1개의 발화로 등록된 성문보다 훨씬 더 성숙한다. 성숙 인자의 다른 예에는 전체 순 음성(net speech)의 지속시간을 포함한다. 예를 들어, 30초 길이의 1 발화로 등록된 화자 성문 모델은 단 2초 길이의 1발화로 등록된 다른 모델보다 더 성숙하다. 또 다른 예에는 오디오의 품질이 포함된다. 예를 들어, 상대적으로 낮은 노이즈의 깨끗한 조건들(높은 SNR, 낮은 T60)에서 수집된 1 발화로 등록된 화자 모델은 시끄럽고 상대적으로 반향하는 조건들(낮은 SNR, 높은 T60)에서 수집된 1 발화로 등록된 모델보다 더 성숙할 것이다.

[0042] 화자 인식 시스템들의 일부 실시예들은 조건 종속적 적응 임계화를 사용한다. 조건 종속적 적응 임계화를 구현함으로써, 시스템은 성숙 결함들을 설명하고 화자들을 인식하거나 인증하도록 훈련된 기계 학습 아키텍처의 원하는 오수락률(또는 원하는 잘못된 식별률)을 충족하도록 정확도를 높인다. 시스템은 성문 성숙도와 성숙 임계값에 기초하여 개별 화자 성문들을 일치시키기 위해 유사도 임계값들을 지속적으로 조정한다. 일부 경우에, 서버는 성숙 인자들 또는 임베딩들과 특정 화자 프로파일의 성문과 관련된 이러한 인자들의 조합에 기초하여 개별 화자 프로파일들에 대한 상이한 유사도 임계값들을 결정할 수 있다. 서버는 관리 디바이스로부터 수신된 관리 구성들에 따라 구성된 대로, 타겟 오수락률 및 성숙 인자들에 따라 주어진 화자에 대한 일련의 유사도 점수 임계값들을 생성하고 업데이트한다. 모델 임베딩(성문)의 경우, 시스템은 성문에 추가된 발화들의 수와 같은, 상이한 수락률 또는 타겟 오정확률 및/또는 성숙 인자 임계값들에 따라 유사도 임계값들을 결정하거나 업데이트한다. 일 예로서, 시스템이 화자-임베딩 모델에 발화들을 추가함에 따라, 시스템이 화자에 대한 더 나은 그림을 갖게 되므로 유사도 임계값이 증가(예를 들어, 화자와 관련된 발화들의 수 증가)할 수 있다. 다른 예로서, 주어진 성문에 대한 유사도 임계값은 서버에 대한 사용자 구성 입력에 따라 구성된 오정확률이 증가할 때 감소할 수 있다.

[0043] 조건 종속적 적응 임계화를 구현하는 시스템은 화자 임베딩 모델을 업데이트하기 위해 수신된 발화들을 사용할 수 있다. 예를 들어, 일부 실시예들에서, 서버는 인바운드 임베딩과 성문을 비교한 다음 특정 화자의 더 높은 임계값과 더 낮은 임계값에 대해 유사도 점수를 평가함으로써, 인바운드 발화로부터 추출된 인바운드 임베딩에 대한 유사도 점수를 생성한 이중 임계화를 사용한다. 유사도 점수가 더 높은 임계값을 초과하면, 서버는 화자를 검증하거나 인증한다. 그런 다음 서버는 성문에 인바운드 임베딩을 추가하고 인바운드 발화를 화자와 연관된 새 발화로서 화자 프로파일에 추가한다. 유사도 점수가 하한 임계값을 초과하면, 서버는 인바운드 화자를 검증하거나 인증한다. 예측 점수가 낮은 임계값을 충족하지만 높은 임계값을 충족하지 않는 상황들에서, 서버는 인바운드 임베딩 및 인바운드 발화를 약한 임베딩 목록에 저장하며, 이는 화자를 검증하기 위해 성문에 충분히 가깝지만 어쩌면 오디오 품질이 좋지 않거나 백그라운드 노이즈로 인해 성문을 업데이트할 만큼 유사하지 않은 임베딩들에 대한 버퍼 또는 격리의 역할을 하는 메모리 위치이다. 시스템은 서버가 성문(기계 학습 아키텍처의 다른 측면들)을 업데이트할 때 성문에 대해 저장된 약한 임베딩 및 발화에 대한 새로운 유사도 점수를 계산하여 저장

된 약한 임베딩 및 발화가 더 높은 임계값을 초과할 만큼 업데이트된 성문과 충분히 유사해져서 모델에 새로운 발화로서 발화를 추가할 수 있는지 여부를 결정한다. 일부 실시예들에서, 서버는 또한 성문을 생성하기 위해 서버에 의해 사용된 하나 이상의 강한 임베딩 목록을 포함한다.

[0044] **비지도(Unsupervised) 클러스터링**

[0045] 음성 생체 인식 시스템은 비지도 클러스터링 방법을 사용하여 한 번에 여러 화자들을 식별할 수 있다. 서버는 유사도 점수들을 계산하기 위한 임의 개수의 클러스터링 알고리즘들 또는 동작들을 실행하여 클러스터들을 생성하고, 성문을 포함하여 임의 개수의 특징들 또는 데이터 유형들을 참조할 수 있다. 클러스터들은 최대 화자들의 임계 수까지 다수의 화자들과 연관되며 화자의 각각의 클러스터와 가장 유사한 발화들에 기초하여 실시간으로 화자들을 식별한다. 예를 들어, 미디어 서비스의 미디어 콘텐트 서버는 가입자 식별자를 가구 또는 가구의 고급 사용자에게 발행한 다음 미디어 데이터베이스에 미리 결정된 명수의 사용자들을 할당한다. 화자 프로파일 데이터베이스는 가입자 식별자와 관련된 사용자들의 수에 따라 하나 이상의 화자 프로파일들을 생성한다. 클러스터링 동작들을 수행하는 서버는 화자 프로파일 또는 미디어 데이터베이스를 참조하여 가입자 식별자에 할당된 사용자들의 수를 결정하고 할당된 사용자들의 수를 화자들의 임계 명수로 사용한다. 인바운드 오디오 신호의 다수의 화자들에 대해 추출된 다수의 임베딩들을 비교하는 것과 같은 클러스터링 동작들에 기초하여, 서버는 유사도 점수들을 생성하고, 가장 근접하게 일치하는 성문들을 식별하고, 가장 가까운 성문들의 유사도 점수를 각 성문들의 유사도 임계값들 또는 기본 유사도 임계값과 비교한다.

[0046] 음성 생체 인식 시스템은 증분 클러스터링(예를 들어, 순차적 클러스터링) 및/또는 조직적 클러스터링(예를 들어, 계층적 클러스터링) 기술들을 사용하여 개별 화자들을 위한 클러스터들을 구축하여 수동적 및 지속적 등록과 인증에 사용될 수 있는 효율적이고 정확한 클러스터링 방법을 보장할 수 있다. 시스템은 특정 기준이 충족되지 않는 한(예를 들어, 예정된 시간인지, 예정된 시간 간격이 경과했는지, 시스템이 이전에 계층적 클러스터링을 사용한 이후로 시스템이 미리 결정된 수의 발화들을 처리했는지, 시스템이 임계값을 초과하는 화자의 수를 식별했는지 등) 증분 클러스터링 동작들을 사용할 수 있으며, 이 겨우 시스템은 조직적 클러스터링 동작들을 실행할 수 있다.

[0047] 예를 들어, 순차적 클러스터링을 사용하기 위해, 음성 생체 인식 시스템은 기존 클러스터 그룹에 대한 새로운 발화에 대한 유사도 점수들을 결정할 수 있다. 시스템은 가장 높은 유사도 점수를 식별하고 유사도 점수가 미리 정의된 임계값을 초과하는지 여부를 결정할 수 있다. 유사도 점수가 임계값을 초과하면, 시스템은 유사도 점수와 관련된 클러스터에 발화를 추가할 수 있다. 그렇지 않으면 시스템이 해당 발화를 첫 번째 발화로 사용하여 새 클러스터를 생성할 수 있다. 시스템은 수집하는 각각의 새로운 발화에 대해 순차적 클러스터링을 구현하여 개별 화자들에 대한 최신 화자 임베딩 모델들을 유지하면서 그렇게 하기 위해 필요한 처리 리소스들을 최소화할 수 있다.

[0048] 예를 들어, 계층적 클러스터링을 사용하기 위해, 음성 생체 인식 시스템은 시스템에 저장된 발화들 각각에 액세스하고 클러스터들 간에 발화를 섞을 수 있다. 시스템은 클러스터들의 발화들 각각을 서로 비교하고 가장 높은 유사도를 갖는 발화들을 함께 클러스터링할 수 있다. 추가로 또는 대안으로, 시스템은 성문들을 서로 비교하여 성문 유사도 임계값과도 일치하는 가장 높은 유사도 점수들을 가진 성문들을 결합한다.

[0049] 각 클러스터링 방법론에는 그 자체의 장점과 단점이 있기 때문에(예를 들어, 증분 클러스터링은 더 빠르지만 덜 정확할 수 있는 반면 조직적 클러스터링은 더 정확하지만 많은 양의 컴퓨터 리소스들이 필요할 수 있음), 시간이 지남에 따라 두 가지 방법론들의 조합을 사용하면 두 방법들의 결함들을 커버하고 시스템이 성숙하고 정확한 화자 임베딩 모델들을 생성하도록 할 수 있다. 시스템은 조직적 클러스터링 동작들과 함께 간헐적으로 증분적 클러스터링 동작들을 실행하여 화자 임베딩 모델들의 정확도를 개선하는 동시에 조직적 클러스터링을 너무 자주 사용하는 것을 피하여 처리 리소스들을 절약할 수 있다. 이 조합은 수동적 및 지속적 등록 및 인증에 적합한 효율적이고 정확한 클러스터링을 보장한다.

[0050] **라벨 정정**

[0051] 재구성 클러스터링 동작들을 사용하는 것은 일련의 라벨 정정 동작들을 구현하기 위해 음성 생체 인식 시스템이 필요할 수 있다. 예를 들어, 재구성 재클러스터링 동작들을 통해 생성된 익명(anonymous) 클러스터들(예를 들어, 새로 생성된 클러스터들, 할당되지 않은 클러스터들)로 라벨들을 정확하게 마이그레이션하기 위해, 시스템은 조직화되지 않은 클러스터들의 클러스터들과 재구성 동작들을 사용하여 조직화된 클러스터들 사이의 쌍별 유사도들을 계산할 수 있다. 시스템은 각각의 이전 클러스터와 새로운 클러스터 사이의 쌍별(pairwise) 유사도

들을 계산하여 유사도 매트릭스를 생성하고 서로 가장 유사한 클러스터들을 매칭 클러스터들로 식별할 수 있다. 시스템은 이전 클러스터들로부터 일치하는 새 클러스터들로 라벨들을 마이그레이션할 수 있다. 시스템은 라벨들과 라벨들에 대한 정보(예를 들어, 콘텐트 선호도) 사이의 연관성들을 저장할 수 있기 때문에, 시스템은 마이그레이션된 라벨들을 통해 이전 클러스터들과 연관되었던 임의의 정보와 새 클러스터들 사이의 연관성들을 생성 및/또는 유지할 수 있다.

[0052] **콘텐트 개인화 및 제어**

[0053] 개인들이 각각의 서비스를 사용하여 콘텐트(예를 들어, 사진 콘텐트, 비디오 콘텐트, 오디오 콘텐트 등) 또는 콘텐트에 대한 권장사항들을 개인들에게 제공함에 따라 개인들에 대한 프로파일들을 생성하거나 제3자 서비스들을 사용하여 생성할 수 있다. 이러한 서비스들은 개인들에 의한 능동적인 입력들로부터 그렇게 할 수 있거나(예를 들어, 사용자들은 그들이 선호하는 콘텐트 유형들을 나타내는 기본 설정을 입력하거나 그들의 연령과 같은 그들 자신에 대한 정보를 제공할 수 있거나) 또는 서비스들은 데이터베이스의 개인들에 대한 프로파일들을 유지하고 사용자들이 오버-더-톱 서비스들을 사용하는 동안 보는 다양한 유형의 콘텐트를 식별할 수 있다. 본원에 설명된 시스템들 및 방법들을 구현함으로써, 시스템은 (예를 들어, 에지(edge) 디바이스를 통해) 오버-더-톱 서비스로부터 수신된 음성 데이터를 사용하여 콘텐트를 보기 위해 서비스를 사용하고 있는 개인들을 식별하고 개인에게 관련 콘텐트를 제공하기 위해 서비스에 개인의 식별자들을 제공할 수 있다.

[0054] 예를 들어, 음성 생체 인식 시스템은 화자로부터 발화들을 수신하고 기계 학습 기술들(예를 들어, 클러스터링 및/또는 신경망 아키텍처)을 사용하여 시스템에 의해 유지된 화자 프로파일 데이터베이스(때로는 "분석 데이터베이스"라고도 함)로부터 화자에 대한 화자 프로파일을 식별할 수 있다. 시스템은 화자 프로파일로부터 화자(또는 화자의 가구, 화자와 관련된 그룹 등)의 하나 이상의 식별자들을 검색하고 화자를 미디어 서비스에 나타낸다. 미디어 서비스는 식별자를 사용사여 서비스가 유지 관리하는 미디어 콘텐트 데이터베이스로부터 식별자와 관련된 프로파일(예를 들어, 소비자 프로파일, 사용자 프로파일)을 식별할 수 있다. 서비스는 식별자와 연관된 프로파일에 기초하여 화자 또는 화자의 에지 디바이스에 콘텐트 및/또는 콘텐트 제안들을 제공할 수 있다. 일부 경우에, 시스템은 화자의 식별자 자체에 기초하여 화자 또는 에지 디바이스에 제공할 콘텐트를 식별할 수 있다. 시스템은 비지도 시스템의 식별자들을 익명(anonymized) 식별자들(예를 들어, 식별자들의 해시된 버전들)로 결정할 수 있기 때문에, 식별자들은 화자들의 익명성을 유지할 수 있으므로 시스템이나 오버-더-톱 서비스가, 식별자들로부터, 시스템 또는 서비스가 콘텐트를 서비스하는 특정 개인들에 대한 개인 식별 정보를 획득할 수 있다.

[0055] 일부 경우에, 음성 생체 인식 시스템은 연령 관련 부모 통제를 위해 화자 프로파일들을 사용할 수 있다. 그렇게 하기 위해, 시스템은 화자의 연령 관련 특성들(예를 들어, 사람이 특정 연령 이상인지 또는 화자의 연령인지 여부)을 나타내는 플래그들과 화자 프로파일들 사이의 연관성들을 저장할 수 있다. 시스템은 사용자 입력을 통해 또는 시스템이 화자의 프로파일을 구축하는 데 사용하는 발화들에 기초하여 자동으로 프로파일들과 관련된 연령 특성들을 획득할 수 있다. 일부 경우에, 시스템은 제3자 서비스로부터 연령 특성들을 수신할 수 있다. 시스템은 개인이 연령 관련 화자 특성을 충족하지 않는 경우 화자에게 제공할 콘텐트를 선택하거나 아니면 화자가 콘텐트를 보는 것을 중지하는 데 사용하기 위해 오버-더-톱 서비스에 연령 관련 특성들을 제공할 수 있다.

[0056] 일부 경우에, 음성 생체 인식 시스템은 본원에 설명된 시스템들 및 방법들을 사용하여 (예를 들어, 개인이 다른 화자의 레코딩을 재생하는 재생 공격에서) 스푸핑된 화자의 화자 프로파일과 관련된 콘텐트를 보기 위해 화자들이 다른 화자들을 스푸핑하는 것을 중지할 수 있다. 예를 들어, 아동은 오버-더-톱 서비스의 연령 관련 제한들을 극복하기 위해 아동의 부모가 말하는 레코딩을 재생할 수 있다. 시스템은 아동이 레코딩을 재생하고 있음을 검출하고, 아동의 부모의 화자 프로파일을 식별하는 대신, 경보를 생성하고/하거나 아동이 자신의 부모를 스푸핑하려고 시도하고 있음을 나타내는 신호를 오버-더-톱 서비스에 전송하여 서비스가 아동에게 연령 제한 콘텐트를 제공하지 못하도록 한다. 따라서 시스템은 개인이 다른 개인을 스푸핑하고 있는지 결정하고 서비스가 인증되지 않은 사용자들에게 콘텐트를 제공하지 못하도록 하는 경고를 생성할 수 있다.

[0057] 재생 공격을 방지하기 위해, 서버는 화자로부터 수신된 추가 인증 데이터를 평가하고 인증 데이터를 추가 암호, 패스프레이즈 또는 기타 필수 정보와 같은 예상된 인증 데이터와 비교할 수 있다.

[0058] 일부 실시예들에서, 서버는 스푸핑 조건들에 대한 다양한 유형의 데이터 및 특징들을 평가함으로써 재생 공격을 방지한다. 서버는 본원에 참조로 통합된, 미국 제10,692,502호로 발행된 미국 출원 번호 제15/910,387호에 설명된 바와 같이, 스푸핑 조건을 나타내는 아티팩트들에 대한 인바운드 오디오 신호를 평가하도록 훈련된 기계 학습 아키텍처 모델을 실행한다. 부모의 발화의 레코딩을 포함하는 재생 공격은, 특정 낮은 레벨 특징들에서, 일

반적으로 실제 라이브 발화들에는 없는 재생 레코딩에서 발견되는 품질들을 포함할 수 있다. 예를 들어, 레코딩된 오디오 샘플은 주파수, 주파수 범위, 동적 전력 범위, 잔향, 특정 주파수 범위의 노이즈 레벨 등과 관련된 오디오 아티팩트들을 일관되게 도입할 수 있으며, 그 아티팩트들 중 적어도 일부는 특수 음성 처리 기술들 및/또는 본원에 설명된 바와 같은 장비를 사용하지 않고는 인지할 수 없을 수 있다. 기계 학습 아키텍처에는 이러한 스푸핑 조건들(예를 들어, 품질들, 아티팩트들)과 실제 인바운드 발화들 사이를 구별하도록 훈련된 스푸핑-검출 분류기가 포함된다. 서버는 잠재적인 스푸핑 조건들을 식별하기 위해 부모에 대한 성문을 참조한다. 예를 들어, 부모는 특정 저레벨 오디오 품질들을 갖는 오디오 신호들을 생성하고 이에 따라 이러한 품질들에 기반한 임베딩들 및 특정하고 일관된 저레벨 특징들에 기반한 부모 성문을 생성하는 특정 디바이스만을 사용하여 발화들을 일관되게 제공할 수 있다. 화자 프로파일은 스푸핑된 액세스와 정품 액세스를 구별하는 데 추후에 사용하기 위해 특정 유형의 저레벨 특징들을 캡처하고 저장할 수 있다. 아동이 인바운드 재생 발화를 입력하면, 서버는 부모의 성문 점수를 생성하는 데 사용된 인바운드 재생 임베딩과 부모 성문의 비교 그리고 인바운드 재생 임베딩과 아티팩트 특징들의 비교에 기초하여 유사도 점수 및/또는 스푸핑 점수를 결정한다. 재생 공격들을 방지하기 위해 서버에 의해 구현될 수 있는 스푸핑 검출의 추가 예들은 본원에 참조로 통합된 미국 출원 제17/192,464호에서 찾을 수 있다.

[0059] **디바이스 개인화 및 제어**

[0060] 일부 디바이스들은 본원에 설명된 것과 같은 화자 프로파일들을 사용하여 에지 디바이스들을 구성 및/또는 사용자 맞춤화할 수 있다. 예를 들어, 시스템은 음성 프로파일들을 자동차에(또는 자동차와 관련된 식별자가 있는 클라우드에서) 로컬로 저장할 수 있으며, 각 음성 프로파일은 자동차의 구성들과 연관된다. 하나의 화자 프로파일과 연관된 개인은 온도, 라디오 볼륨, 창 설정 등 중 하나 이상과 연관될 수 있고, 다른 개인은 다른 설정들과 연관될 수 있다. 시스템은 다수의 개인들에 대한 프로파일들을 저장할 수 있으므로 시스템이 발화를 수집하고 발화들에 기초하여 화자 프로파일들을 식별할 때, 시스템은 디바이스의 다른 애플리케이션들과 통신하여 화자 프로파일의 설정들에 기초하여 디바이스의 구성들을 자동으로 조정할 수 있다.

[0061] **예시적인 시스템 컴포넌트들**

[0062] **미디어 콘텐트 시스템을 위한 오디오 처리**

[0063] 설명과 이해의 편의를 위해, 본원에 설명된 실시예들은 콘텐트 전달 시스템의 맥락에서 오디오 처리 및 사용자 데이터 분석을 사용하는 컴퓨팅 시스템들을 설명한다. 그러나 실시예들은 이러한 구현들에 제한되는 것은 아니며, 개별 또는 다중-화자 음성 생체 인식을 처리하기 위해, 예를 들어 수동적 또는 능동적 화자 식별, 또는 지속적 또는 정적 화자 식별의 이점을 얻을 수 있는 임의 개수의 시스템들에 대해 본원에 설명된 프로세스들을 사용할 수 있다. 오디오 입력들에서 화자들을 수신, 처리 및 식별하는 거의 모든 시스템은 본원에 설명된 다중-화자 식별 또는 다중-화자 음성 생체 인식의 인증을 위한 시스템들 및 프로세스들을 구현할 수 있다. 본원에서 오디오 처리 및 데이터 분석을 사용할 수 있는 시스템들의 비제한적인 예들에는, 그 중에서도, IoT 디바이스들(때때로 에지 디바이스들이라고도 함)(예를 들어, 스마트 기기들, 차량들), 콜 센터들 및 유사한 헬프 데스크들 또는 서비스 센터들, 보안 인증 시스템들 또는 서비스들(예를 들어, 사무실 또는 가정 보안), 및 감시 또는 인텔리전스 시스템이 포함된다.

[0064] 게다가, 본원의 실시예들은 화자들을 특정 알려진 또는 알 수 없는 사용자들로 식별하기 위해 오디오 처리 동작들을 사용한다. 그러나 실시예들은 음성 생체 인식으로만 제한되는 것은 아니며, 화자들을 특정 사용자들로 식별하기 위해 임의 개수의 추가 유형의 생체 인식들을 수집 및 처리할 수 있다. 실시예들이 수집 및 처리할 수 있는 추가 유형의 생체 인식들의 비제한적인 예들에는 모니터링되는 시스템(예를 들어, 키 누르기, 메뉴 액세스, 콘텐트 선택, 입력 또는 선택 속도) 또는 임의의 생체 인식 정보의 조합에 액세스할 경우의 눈 스캔(예를 들어, 망막 또는 홍채 인식), 얼굴(예를 들어, 얼굴 인식), 지문 또는 손자국(예를 들어, 지문 인식), 사용자 행동(behaviors)(예를 들어, "행동 인쇄(behavior prints)")이 포함된다.

[0065] 도 1은 오디오 처리 기계 학습 동작들을 사용하는 시스템(100)의 컴포넌트들을 도시한다. 시스템(100)은 분석 시스템(101), 콘텐트 시스템(110), 및 최종 사용자 디바이스들(114)을 포함한다. 분석 시스템(101)은 분석 서버(102), 분석 데이터베이스(104), 및 관리 디바이스(103)를 포함한다. 콘텐트 시스템(110)은 콘텐트 서버들(111) 및 콘텐트 데이터베이스들(112)을 포함한다. 실시예들은 추가 또는 대안적인 컴포넌트들을 포함하거나 도 1의 컴포넌트들로부터 특정 컴포넌트들을 생략할 수 있으며, 여전히 본 개시의 범위에 속한다. 예를 들어, 다수의 콘텐트 시스템들(110)을 포함하거나 분석 시스템(101)이 다수의 분석 서버들(102)을 갖는 것이 일반적일 수 있다. 실시예들은 본원에 설명된 다양한 특징들 및 태스크들을 수행할 수 있는 임의 개수의 디바이스들을 포함하

거나 아니면 구현할 수 있다. 예를 들어, 도 1은 분석 데이터베이스(104)와 별개의 컴퓨팅 디바이스로서의 분석 서버(102)를 도시한다. 일부 실시예들에서, 분석 데이터베이스(104)는 분석 서버(102)와 통합된다.

[0066] 시스템(100)의 다양한 컴포넌트들은 하나 이상의 공용 또는 사설 네트워크들의 하드웨어 및 소프트웨어 컴포넌트들을 통해 서로 상호연결된다. 이러한 네트워크들의 비제한적인 예들에는 근거리 통신망(Local Area Network; LAN), 무선 근거리 통신망(Wireless Local Area Network; WLAN), 대도시 통신망(Metropolitan Area Network; MAN), 광역 통신망(Wide Area Network; WAN) 및 인터넷이 포함될 수 있다. 네트워크를 통한 통신은 전송 제어 프로토콜 및 인터넷 프로토콜(Transmission Control Protocol and Internet Protocol; TCP/IP), 사용자 데이터그램 프로토콜(User Datagram Protocol; UDP), IEEE 통신 프로토콜들과 같은 다양한 통신 프로토콜들에 따라 수행될 수 있다.

[0067] **개요 및 인프라스트럭처**

[0068] 분석 서비스는 미디어 콘텐트 서비스 또는 엔터프라이즈 콜 센터들과 같은 고객들에게 오디오 처리를 위한 컴퓨팅 서비스들 및 최종 사용자로부터 수신된 데이터 분석을 제공한다. 분석 서비스들의 비제한적인 예들에는 사용자 식별, 화자 인식(예를 들어, 화자 분리(diarization)), 사용자 인증 및 최종 사용자들에 대한 데이터 분석이 포함된다. 분석 서비스는 하나 이상의 미디어 콘텐트 서비스들(예를 들어, Netflix®, TiVO®)의 하나 이상의 콘텐트 시스템들(110)에 분석 서비스들을 호스팅하고 제공하도록 구성된 다양한 하드웨어, 소프트웨어 및 네트워킹 컴포넌트들을 포함하는 분석 시스템(101)을 동작시킨다.

[0069] 콘텐트 미디어 서비스는 콘텐트 시스템(110)을 동작시키며, 미디어 콘텐트를 최종 사용자 디바이스들(114)에 제공하는, 오버-더-톱(OTT) 또는 디지털 스트리밍 서비스들과 같은 클라우드 기반 미디어 서비스들을 호스팅하도록 구성된 하드웨어, 소프트웨어 및 네트워킹 컴포넌트들을 포함한다. 콘텐트 시스템(110)은 콘텐트 추천, 콘텐트 제한(예를 들어, 부모 컨트롤), 및/또는 광고와 같은 최종 사용자들을 위한 개인화된 콘텐트를 식별하고 제공한다.

[0070] 동작 시, (콘텐트 시스템(110)의) 제공자 서버(111)는 최종 사용자 디바이스들(114)로부터 다양한 유형의 입력 데이터를 수신하고 입력 데이터를 (분석 시스템(101)의) 분석 서버(102)로 포워딩한다. 분석 서버(102)는 제공자 서버(111)로부터 포워딩된 입력 데이터를 사용하여 본원에 설명된 다양한 분석 프로세스들을 실행한 다음, 분석 프로세스들에서 생성된 다양한 출력들을 제공자 서버(111)로 전송한다. 제공자 서버(111)는, 콘텐트 개인화를 위한 다른 유형의 정보 중에서, 분석 서버(102)로부터 수신된 출력들을 사용하여, 예를 들어 사용자 동작들 또는 행동들(예를 들어, 시청 습관), 최종 사용자 디바이스들(114)과 콘텐트 시스템(110) 사이의 상호작용들, 사용자들의 특성들(예를 들어, 연령), 및 사용자 ID들을 식별 및 생성한다.

[0071] 콘텐트 시스템(110) 또는 분석 시스템(101)은 일반적으로 예를 들어 가입 정보(예를 들어, 가입자 식별자) 또는 사용자 자격 증명(예를 들어, 사용자 이름, 비밀번호)에 기초하여 사용자를 식별할 수 있지만, 본원에 설명된 분석 시스템(101)은 최종 사용자 디바이스들(114)에 의해 캡처된 사용자 입력들 및 음성 발화들에 기초하여 (콘텐트 시스템(110)을 대신하여) 사용자를 추가로 또는 대안으로 식별한다.

[0072] 일부 경우에, 최종 사용자 디바이스(114)는 사용자 입력 데이터를 능동적으로 캡처하며, 여기서 최종 사용자는 최종 사용자 디바이스(114)와 능동적으로 상호작용(예를 들어, "웨이크(wake)" 단어 말하기, 버튼 누르기, 제스처 만들기)한다. 일부 경우에, 최종 사용자 디바이스(114)는 사용자 입력 데이터를 수동적으로 캡처하며, 여기서 최종 사용자는 최종 사용자 디바이스(114)와 수동적으로 상호작용한다(예를 들어, 다른 사용자에게 말을 하는 경우, 최종 사용자 디바이스(114)는 사용자의 적극적인 조치 없이 자동으로 발화를 캡처함). 다양한 유형의 입력들은 사용자들이 최종 사용자 디바이스(114)의 마이크에 의해 캡처된 사운드 또는 오디오 데이터 또는 최종 사용자 디바이스(114)에 의해 제시된 사용자 인터페이스를 통해 입력된 사용자 입력들과 같은 최종 사용자 디바이스들(114)과 상호 작용하는 방식들을 나타낸다. 캡처된 사운드에는 백그라운드 노이즈(예를 들어, 주변 소음) 및/또는 한 명 이상의 화자-사용자들의 발화들이 포함된다. 추가로 또는 대안으로, 사용자 입력들은 최종 사용자 디바이스(114)에 의해 캡처되거나 이에 업로드된 사용자들의 비디오들(또는 이미지들)(예를 들어, 얼굴 표정들, 제스처들)을 포함할 수 있다. 사용자 인터페이스에 대한 사용자 입력들은 물리적 또는 그래픽 사용자 인터페이스에 대한 인터페이스 입력들, 예컨대 디바이스를 가로질러 스와이프하는 터치 입력, 제스처로 디바이스 사용, 디바이스 상의 버튼 누르기(예를 들어, 키패드 이중 톤 멀티- 주파수(dual-tone multi-frequency; DTMF) 톤), 텍스트 입력, 지문과 같은 생체 인식 정보 캡처 등을 포함할 수 있다.

[0073] 콘텐트 서버(111)는 사용자 입력 데이터로서 최종 사용자 디바이스들(114)로부터 사용자 입력들을 수신한다. 콘

텐트 서버(111)는 다양한 형태의 메타데이터 식별 또는 추출, 사용자 입력 데이터를 사용하여 하나 이상의 인증 동작 수행, 특정 유형의 데이터를 익명화 또는 난독화(예를 들어, 하나 이상의 식별자들의 해시 생성)와 같은 사용자 입력 데이터에 대한 다양한 처리 동작들을 수행한다. 콘텐트 서버(111)는 사용자 입력 데이터를 분석 서버(102)로 전송하기 전에 사용자 입력 데이터를 변환, 수정 및/또는 강화할 수 있다. 사용자 입력 데이터의 데이터 유형의 비제한적인 예들에는, 그 중에서도, 오디오 신호들, 사용자 인터페이스 입력들(예를 들어, 사용자 요청들, 사용자 명령어들), 인증 입력들(예를 들어, 사용자 자격 증명, 생체 인식), 다양한 식별자들(예를 들어, 가입자 식별자들, 사용자 식별자들) 및 다양한 유형의 통신 메타데이터가 포함된다.

[0074] 분석 시스템(101)은 제공자 서버(111)로부터 오디오 신호를 데이터 파일 또는 데이터 스트림으로서 임의 개수의 기계 판독 가능 데이터 형식들(예를 들어, WAV, MP3, MP4, MPEG, JPG, TIF, PNG, MWV)로 수신한다. 오디오 신호에는 백그라운드 노이즈 외에 음성이 포함될 수 있다. 일부 구성들에서, 분석 시스템(101)이 오디오 신호를 수신하기 전에, 콘텐트 시스템(110)은 오디오 신호에 메타데이터 또는 화자-독립 특성들과 같은 상호작용에 대한 정보를 태깅할 수 있다. 이러한 정보의 비제한적인 예들에는 상호작용 시간, 상호작용 날짜, 최종 사용자 디바이스(114)의 유형, 마이크 유형, 상호작용 장소(예를 들어, 침실, 거실, 식당), 상호작용과 연관된 특정 최종 사용자 디바이스(114)(예를 들어, 특정 스마트 TV(114a), 가상 어시스턴트(114d)), 상호작용과 연관된 가입자 식별자, 상호작용과 연관된 고유 식별자(예를 들어, 자동 번호 식별자) 등이 포함될 수 있다.

[0075] 분석 서버(102)와 같은 분석 시스템(101)의 컴포넌트들은 사용자 식별 서비스들(또는 인증 서비스들)을 제공하기 위해, 성문을 생성하고, 성문을 업데이트하고, 유사도 점수를 예측하고, 오디오 신호에서 화자들을 식별(또는 인증)하고, 성문에 다시 레이블링한다. 분석 시스템(101)은 오디오 신호의 화자와 관련된 유사도 점수(또는 라벨)를 분석 시스템(101)의 고객들에게 제공할 수 있다. 예를 들어, 분석 시스템(101)은 상호작용과 연관된 오디오 신호에 기초하여 화자 식별자를 콘텐트 시스템(110)에 전송할 수 있다.

[0076] **콘텐트 시스템**

[0077] 콘텐트 시스템(110)은 미디어 데이터를 가입자들의 최종 사용자 디바이스들(114)로 전송하거나 스트리밍한다. 가입자는 콘텐트 시스템(110)의 고객을 나타내지만, 또한 하나 이상의 사용자들의 집합을 나타낼 수도 있다. 예를 들어, 가입자는 최종 사용자 디바이스들(114)을 사용하여 콘텐트 시스템(110)의 서비스들에 액세스하는 가구(household) 및 가족 및/또는 게스트의 구성원들을 나타낼 수 있다. 가족 중 적어도 한 구성원이 콘텐트 시스템(110)에 가입자로 등록한 경우 가족은 미디어 콘텐트에 액세스할 수 있다. 제공자 서버(111)는 최종 사용자 디바이스들(114)과의 다양한 사용자 상호작용들에 기초하여 사용자들(가입자들 및 가입자들의 게스트들을 포함)을 위해 최종 사용자 디바이스들(114)에 미디어 콘텐트를 전송한다.

[0078] 콘텐트 시스템(110)의 컴포넌트들은 최종 사용자 디바이스(114)로부터 수신된 다양한 유형의 입력 데이터(예를 들어, 메타데이터)를 캡처하고, 콘텐트 시스템(110)은 하나 이상의 네트워크들을 통해 분석 서버(102)에 포워딩된 입력 데이터에 추가한다. 콘텐트 서버들(111)과 같은 콘텐트 시스템(110)의 컴퓨팅 디바이스들은 입력 데이터를 분석 시스템(101)에 포워딩하기 전에 가입자 식별자에 인바운드 입력 데이터(오디오 신호 포함)를 할당한다. 가입자 식별자는 콘텐트 시스템(110)의 가입자 또는 고객을 나타내는 데이터 값, 태그 또는 다른 형태의 데이터이다. 가입자 식별자는 콘텐트 시스템(110)이 특정 가입자로부터 특정 입력과 함께 수신된 오디오 신호를 포함하여는 특정 가입자와 다양한 유형의 데이터를 연관시킨다.

[0079] 가입자는 사용자들의 집합에 공통적인 가입자 식별자를 사용하여 콘텐트 시스템(110)의 서비스들에 액세스하는 사용자들의 집합(예를 들어, 가구)을 정의하거나 아니면 연관시킨다. 이러한 경우에, 콘텐트 데이터베이스(112)는 때때로 "화자 프로파일(speaker profiles)"이라고 하는 사용자들(화자들) 각각에 대한 데이터 레코드들을 포함한다. 동작 시, 분석 시스템(101)은 입력 데이터에 대한 가입자 식별자 및 오디오 신호를 수신하지만, 프라이버시를 위해 분석 시스템(101)은 개인 식별 정보를 포함할 필요가 없다. 예를 들어, 콘텐트 서버(111)는 익명화된 버전의 가입자 식별자 또는 화자 식별자를 생성하여 분석 시스템(101)에 전송함으로써, 가입자의 개인 정보가 분석 시스템(101)으로 직접 전송되는(또는 아니면 확인할 수 있는) 것을 방지함으로써 가입자 또는 특정 사용자들의 개인 정보를 보호한다.

[0080] 콘텐트 시스템(110)은 콘텐트 데이터베이스(112)에 미디어 및 가입자 정보를 저장하여, 콘텐트 시스템(110)이 가입자 식별자와 연관된 가입된 사용자들(예를 들어, 가입자의 계정)을 식별하도록 한다. 예를 들어, 콘텐트 데이터베이스(112)는 가입자 식별자, 오디오 신호(분석 시스템(101)로부터 수신된 화자 식별자들, 사용자 특성들, 화자-독립 특성들 및 기타 메타데이터 포함) 및 가입자 계정을 포함하는 가입자 데이터 레코드들을 룩업 테이블(lookup table)에 저장한다.

[0081] 콘텐트 시스템(110)은 미리 구성된 트리거링 조건들에 따라 오디오 신호(가입자 식별자를 가짐)를 분석 시스템(101)에 포워딩한다. 예를 들어, 콘텐트 시스템(110)은 최종 사용자 디바이스(114)로부터 오디오 신호를 수신하고 오디오 신호를 분석 시스템(111)으로 포워딩한다. 오디오 신호는 예를 들어, 사용자가 콘텐트 시스템(110)에 미디어 콘텐트(또는 다른 서비스들)를 요청하는 하나 이상의 발화들을 포함한다. 그 콘텐트 시스템(110)은 실시간으로 사용자 세트에 대해 사용자를 식별하기 위해 분석 시스템(101)에 요청과 연관된 오디오 신호를 포워딩할 수 있다. 분석 시스템(101)은 화자 정보(예를 들어, 화자 특성들, 화자 식별자들, 화자-독립 특성들, 메타데이터)를 콘텐트 시스템(110)에 전송할 수 있다. 콘텐트 시스템(110)은 사용자 정보를 사용하여 개인화된 콘텐트로 화자 요청에 응답할 수 있다. 일부 경우에, 콘텐트 서버(111)는 분석 서버(102) 또는 관리 디바이스(103)와 같은 시스템(100)의 다른 디바이스로부터 수신된 명령어들 또는 쿼리들에 응답하여 오디오 신호를 분석 시스템(101)에 포워딩할 수 있다.

[0082] **콘텐트 서버**

[0083] 일부 실시예들에서, 콘텐트 서버(111)는 오디오 신호에서 음성을 식별하고, 오디오 신호를 한 형식에서 다른 형식의 오디오 신호로 변환(예를 들어, 미디어 파일들을 WAV 파일 형식에서 MP3 파일 형식으로 변환)하고, 오디오 신호를 전처리하고, 오디오 신호를 익명화하고(예를 들어, 해시 식별자를 오디오 데이터와 연관시키고), 오디오 신호에서 화자들과 관련된 생체 특징들을 추출하는 등을 하기 위한 소프트웨어 프로세스들 및 서비스들을 호스팅하고 실행할 수 있다. 예를 들어, 콘텐트 서버(111)는 오디오 신호에서 오디오 이벤트들을 검출하도록 구성된다. 콘텐트 서버(111)는 또한 오디오 신호의 콘텐트(예를 들어, 콘텐트 소비에 대한 사용자 요청)를 캡처하기 위해 오디오 신호에 대해 자동 음성 인식(ASR)을 수행하도록 구성될 수 있다.

[0084] 콘텐트 서버(111)는 최종 사용자 디바이스(114)와 능동적으로 상호작용하는 사용자에게 콘텐트를 제공한다. 예를 들어, 사용자는 최종 사용자 디바이스(114)에 말할 수 있다. 콘텐트 서버(111)는 또한 최종 사용자 디바이스(114)와 수동적으로 상호작용하는 사용자에게(예를 들어, 최종 사용자 디바이스(114)에 대한 미리 결정된 근접성 내에서 말하는 사용자들에게) 콘텐트를 제공할 수 있다. 콘텐트 서버(111)는 분석 서버(102)에 의해 표시된 바와 같이 화자들의 ID에 기초하여 텔레비전 쇼 또는 광고의 추천을 포함하는 사용자 인터페이스 데이터 또는 콘텐트(예를 들어, 컴퓨터 파일들, 데이터 스트림)를 화자들에게 전송할 수 있다.

[0085] **콘텐트 데이터베이스**

[0086] 콘텐트 시스템(110)의 콘텐트 데이터베이스(112)는 최종 사용자 디바이스들(114)로 스트리밍하기 위한 가입자 데이터, 화자 프로파일들, 및 미디어 콘텐트를 포함하는 다양한 유형의 데이터 레코드들을 저장한다. 예를 들어, 콘텐트 데이터베이스(112)는 콘텐트 라이브러리를 저장할 수 있다. 콘텐트 데이터베이스(112)는 또한 오디오 신호, 화자 식별자들, 화자 특성들, 화자-독립 특성들, 및 최종 사용자 디바이스들(114) 또는 분석 서버(102)로부터 수신된 상호작용과 연관된 다른 메타데이터를 저장할 수 있다.

[0087] 콘텐트 데이터베이스(112)는 또한 계정의 소유자 또는 가구, 계정의 허용된 사용자 수, 계정과 관련된 허용된 디바이스들의 수, 계정과 관련된 현재 사용자들의 수, 계정과 연결된 현재 디바이스들의 수, 허용되는 지리적 운영 영역(예를 들어, 계정이 일부 국가에서는 금지되고 다른 국가에서는 허용될 수 있음), 구매 옵션들(예를 들어, 모든 구매 전에 비밀번호 요구), 결제 정보(예를 들어, 신용카드 정보, 청구서 주소, 배송 주소), 가입자와 관련된 식별자들(예를 들어, 가입자 식별자, 가구 식별자), 가입자 식별자와 연결된 화자 식별자들, 익명화 정보(예를 들어, 해시 함수, 암호화 키) 등과 같은 가입자 정보를 저장할 수 있다.

[0088] 콘텐트 데이터베이스(112)의 가입자 프로파일 또는 사용자 프로파일은 시청 이력, 화자 정보(예를 들어, 이름, 연령, 생일, 성별, 종교), 보안 자격 증명(예를 들어, 로그인 자격 증명, 생체 인식), 및 특정 화자들에 대한 선호도를 저장할 수 있다. 선호도의 비제한적인 예들에는 화자가 역사적으로 보거나, 좋아하거나, 북마크한 콘텐트, 또는 아니면 사용자가 관심을 표시한 콘텐트 등이 포함될 수 있다. 동작 시, 콘텐트 서버(111)는 로그인 자격 증명 또는 메타데이터에 기초하여 또는 분석 서버(102)에 의해 결정된 바와 같이 특정 화자를 식별하고, 선호도에 기초하여 화자에 대한 특정 미디어를 결정한다. 가입자 프로파일은 하나 이상의 화자 프로파일들을 포함하거나 이와 연관된다. 일부 경우에, 콘텐트 데이터베이스(112)의 화자 프로파일은 분석 데이터베이스(104)(때때로 화자 데이터베이스라고도 함)에 저장된 화자 프로파일에 대응한다.

[0089] 일부 구현들에서, 가입자 또는 화자 프로파일은 콘텐트 서버(111)에게 특정 미디어가 특정 화자들에게 전달되는 것을 금지하도록 지시하는 콘텐트 제한 또는 제어(예를 들어, 미성년자 사용자를 위한 부모 통제)를 포함한다. 콘텐트 제한(가입자 또는 사용자 프로파일들에서)은 콘텐트 데이터베이스(112)의 미디어 데이터 레코드들에 저

장된 연령별 등급(예를 들어, R, PG-13, TV-MA) 또는 콘텐트 특성들에 해당한다. 콘텐트 특성들은 부모/재량 권고 또는 담배/약물 사용, 섬광등, 과도한 노출 및 폭력과 같은 극단적이거나 불쾌한 유형의 콘텐트를 나타내는 데이터 값들이다. 콘텐트 특성들은 콘텐트 데이터베이스(112) 또는 분석 데이터베이스(104)에 사용자 또는 화자 프로파일 정보로서 저장된 사용자 또는 화자 특성들에 대응한다.

[0090] 가입자 프로파일의 적어도 하나의 사용자 프로파일이 전체 가입자 프로파일(예를 들어, 가구) 또는 특정 사용자들(예를 들어, 아동 화자)에 대한 콘텐트 제한을 구성할 수 있는 권한이 있는 파워 사용자 프로파일(예를 들어, 부모의 화자 프로파일)로 지정된다. 파워 사용자는 다양한 콘텐트 제한 구성들을 구성 인터페이스에 입력하기 위해 최종 사용자 디바이스(114)의 사용자 인터페이스를 동작시킨다. 콘텐트 제한 구성들은 특정 사용자 프로파일 식별자(예를 들어, 사용자 이름), 사용자 연령 또는 특정 콘텐트/화자 특성들에 따라 콘텐트 제한을 표시 및 구성한다. 콘텐트 서버(111)는 일반적으로 사용자의 사용자 식별자를 참조함으로써 특정 사용자가 콘텐트 서비스에 로그인했다고 결정할 수 있다. 콘텐트 서버(111)는 로그인 자격 증명(사용자 식별자 포함)에서 사용자 식별자를 수신하거나 분석 서버(102)로부터 반환된 화자 식별자를 사용하여 콘텐트 데이터베이스(112)에서 사용자 식별자를 식별한다. 콘텐트 서버(111)는 사용자 식별자를 사용하여 콘텐트 데이터베이스(112)에 쿼리하고 특정 사용자에게 할당된 권한의 레벨 및 임의의 대응하는 콘텐트 제한을 결정한다.

[0091] 일부 실시예들에서, 구성 제한들은 추가 인증 확인 입력(예를 들어, PIN, 비밀번호) 또는 생체 인식 입력(예를 들어, 지문, 성문)과 같은 콘텐트 제한을 비활성화하거나 업데이트하는 수단을 더 나타낸다. 동작 시, 분석 서버(102)가 입력 데이터(예를 들어, 오디오 신호)로부터 특정 사용자(예를 들어, 화자)를 식별한 후, 콘텐트 서버(111)는 콘텐트 데이터베이스(112)를 참조하여 사용자에게 할당된 콘텐트 제한을 식별하고 미디어 데이터레코드들에 저장된 연령 적정 등급들 또는 콘텐트 특성들에 따라 사용자에게 요청, 쿼리 또는 제시되는 콘텐트에 대한 콘텐트 제한을 적용한다.

[0092] **최종 사용자 디바이스들**

[0093] 최종 사용자 디바이스들(114)은 사용자가 임의의 오디오 또는 시각적 인터페이스를 제어하거나 아니면 콘텐트 서비스들을 동작시키기 위해 운용하는 임의의 디바이스일 수 있다. 최종 사용자 디바이스들(114)의 비제한적인 예들에는 스마트 TV(114a), 리모트 컨트롤러(114b), 셋톱 박스(114c), 또는 가상 어시스턴트(114d) 또는 다른 에지 디바이스 또는 모바일 컴퓨팅 디바이스와 같은 IoT 디바이스가 포함될 수 있다. 스마트 TV(114a)는 인터넷 네트워크와 같은 네트워크에 연결되고 마이크 및/또는 카메라로 구성된 TV일 수 있다. TV 리모트 컨트롤러(114b)는 TV에 디스플레이되는 콘텐트를 제어하는 컨트롤러일 수 있다. 셋톱 박스(114c)(예를 들어, 케이블 박스, Slingbox, AppleTV, Roku Streaming Stick, TiVo Stream, Amazon Fire)는 프로세서 및 비일시적 저장 매체를 포함하고 본원에 설명된 다양한 프로세스들을 수행하도록 구성된 임의의 미디어 스트리밍 및/또는 저장 디바이스를 포함한다. 셋톱 박스(114c)는 다양한 유형의 콘텐트, 사용자 및 디바이스 정보를 업로드 및 다운로드하기 위해 하나 이상의 네트워크들을 통해 미디어 콘텐트 시스템(110)과 통신할 수 있다. IoT 디바이스는 VoIP(Voice-over-IP) 텔레통신 또는 기타 네트워크 통신(예를 들어, 셀룰러, 인터넷)을 구현하도록 구성된 텔레통신 지향 디바이스(예를 들어, 휴대폰) 또는 컴퓨팅 디바이스일 수 있다. IoT 디바이스는 TCP/IP 네트워크 또는 기타 컴퓨팅 네트워크 채널을 통한 데이터 스트리밍을 위해 구성된 하드웨어 및 소프트웨어 컴포넌트들을 포함한다. 개인 어시스턴트(114d)는 가상 어시스턴트 디바이스(예를 들어, Alexa®, Google Home®), 스마트 기기, 자동차, 또는 소프트웨어 애플리케이션들을 실행하고/하거나 음성 인터페이스 동작들을 수행할 수 있는 기타 스마트 디바이스일 수 있다.

[0094] 최종 사용자 디바이스들(114)은 페어링(paired)되거나 아니면 네트워킹된 디바이스의 통신 특징들을 사용할 수 있는 프로세서 및/또는 소프트웨어를 포함한다. 최종 사용자 디바이스들(114)은 마이크, 가속도계, 자이로스코프, 카메라, 지문 스캐너, 상호작용 버튼들(예컨대 방향 버튼, 숫자 버튼), 조이스틱, 또는 임의의 조합 등으로 구성될 수 있다. 최종 사용자 디바이스들(114)은 사운드(예를 들어, 음성 발화, 주변 소음)를 감지하고 전기적 오디오 신호들로 변환하기 위한 하드웨어(예를 들어, 마이크) 및/또는 소프트웨어(예를 들어, 코덱)를 포함할 수 있다. 콘텐트 시스템(110)은 콘텐트 데이터베이스(112)에 오디오 신호를 수집 및 저장할 수 있지만, 분석 시스템(101)은 일반적으로 오디오 신호를 저장하는 것을 회피하거나 메모리에 저장된 임의의 오디오 신호들을 퍼지(purge)한다.

[0095] 콘텐트 서버(111)는 최종 사용자 디바이스들(114)로부터 입력 데이터를 수신하고, 입력 데이터의 오디오 신호를 변환하거나, 가입자 식별자를 식별 및 입력 데이터와 연관시키거나, 오디오 신호의 다양한 특징들을 콘텐트 데이터베이스(112)에 저장하는 것과 같은 다양한 전처리 동작들을 수행할 수 있다. 그런 다음 콘텐트 서버(111)는

오디오 신호(및 가입자 식별자)를 하나 이상의 네트워크들을 통해 (분석 시스템(101)의) 분석 서버(102)에 포워딩한다.

[0096] 콘텐트 서버(111)는 오디오 신호를 미리 결정된 양의 시간 동안 버퍼 또는 캐시 메모리와 같은 비일시적 기계 판독 가능 저장 매체에 일시적으로 저장한다. 추가로 또는 대안으로, 콘텐트 서버(111)는 오디오 신호를 콘텐트 데이터베이스(112)에 저장한다.

[0097] 콘텐트 서버(111)는 입력 데이터(예를 들어, 오디오 신호, 메타데이터)를 분석 시스템(101)의 컴퓨팅 디바이스들(예를 들어, 분석 서버(102))로 포워딩한다. 일부 구성들에서, 콘텐트 서버(111)는 미리 결정된 간격과 같은 미리 구성된 트리거 조건들에 따라 또는 콘텐트 서버(111)가 최종 사용자 디바이스(114)로부터 입력 오디오 신호를 수신하는 것에 응답하여 입력 데이터를 분석 서버(102)에 전송한다.

[0098] 일부 실시예들에서, 콘텐트 서버(111) 또는 최종 사용자 디바이스(114)는 최종 사용자 디바이스(114)가 사용자로부터 능동적 입력(예를 들어, 웨이크 단어)을 검출하기 전에도 사운드 레코딩들을 지속적으로 캡처하고 저장한다. 예를 들어, 한 그룹의 사람들이 무엇을 볼 것인지 논의할 때, 최종 사용자 디바이스(114)는 사운드를 캡처한다. 그룹이 시청할 특정 쇼(show)를 결정한 후, 그룹은 일반적으로 침묵하고 한 사람이 웨이크 단어를 말하고 그룹이 시청하기로 결정한 쇼를 발표한다. 그룹이 침묵했기 때문에, 최종 사용자 디바이스(114) 및 콘텐트 서버(111)는 웨이크 단어 이전의 일정 시간 기간 동안 그룹 토론의 사운드를 캡처하고 저장하는 것으로부터 이익을 얻는다. 이러한 방식으로, 콘텐트 서버(111) 및 분석 서버(102)는 능동적 입력 전에 일정 양의 시간으로부터 오디오 신호들을 검토한다. 이러한 실시예들에서, 사용자가 사운드를 능동적으로 캡처하기 위해 최종 사용자 디바이스(114)를 동작시킬 때, 콘텐트 시스템(110)은 사용자가 최종 사용자 디바이스(114)를 능동적으로 동작시키기 전에 최종 사용자 디바이스(114)에 의해 캡처된 저장된 오디오 신호들을 검색한다. 콘텐트 시스템은 트리거 조건과 관련된 오디오 신호들과 검색된 오디오 신호들 모두를 분석 서버(102)에 포워딩한다.

[0099] 분석 서버(102)는 수신된 오디오 신호들 및 일부 실시예들에서는 추가 유형의 데이터(예를 들어, 가입자 식별자, 사용자 자격 증명, 메타데이터)를 사용하여 화자 식별자를 결정한다. 분석 서버(101)는 화자 식별자(및 화자 특성들, 화자-독립 특성들 및 기타 메타데이터)를 콘텐트 서버(111)에 전송한다. 콘텐트 서버(111)는 화자 식별자를 가입자 식별자 및/또는 특정 화자에 관한 정보에 매핑하여 콘텐트 서버(111)가 화자에게 전달할 필요가 있는 콘텐트를 결정한다.

[0100] **분석 서버**

[0101] 분석 서버(102)는 하나 이상의 프로세서들 및 소프트웨어를 포함하고, 본원에 설명된 다양한 프로세스들 및 태스크들을 수행할 수 있는 임의의 컴퓨팅 디바이스일 수 있다. 분석 서버(102)는 데이터베이스들(112 및 104)을 호스팅하거나 이와 통신할 수 있고, 콘텐트 시스템(110)으로부터 오디오 신호들, 화자-독립 특성들, 및 가입자 식별자를 수신할 수 있다. 도 1은 단일 분석 서버(102)를 도시하지만, 분석 서버(102)는 임의 개수의 컴퓨팅 디바이스들을 포함할 수 있다. 일부 구성들에서, 분석 서버(102)는 클라우드 컴퓨팅 또는 가상 머신 구성에서 동작하는 임의 개수의 컴퓨팅 디바이스들을 포함할 수 있다. 일부 실시예들에서, 콘텐트 시스템(110)의 컴퓨팅 디바이스들(예를 들어, 콘텐트 서버(111))은 분석 서버(102)의 기능들을 부분적으로 또는 전체적으로 수행한다.

[0102] 분석 서버(102)는 예를 들어, 콘텐트 서버(111)로부터 입력된 오디오 신호(예를 들어, 화자 발화의 오디오 레코딩, 가입자 식별자, 사용자 식별자, 메타데이터)를 수집하고, 분석 데이터베이스(104)에 쿼리하고, 오디오 데이터에 다양한 기계 학습 동작들을 적용한다. 기계 학습 알고리즘들은 오디오 이벤트 검출, 임베딩 추출, 등록된 성문 생성 또는 업데이트, 및 오디오 신호에서 발화가 있는 한 명 이상의 사용자들의 식별/인증과 같은, 본원에 설명된 다양한 동작들을 수행하기 위한 임의 개수의 기술들 또는 알고리즘들(예를 들어, 가우스 매트릭스 모델(Gaussian Matrix Model; GMM)들, 신경망들)을 구현한다.

[0103] 분석 서버(102)는 분석 데이터베이스(104)에 저장된 화자 프로파일들을 쿼리하여 오디오 신호들에서 알려진 또는 새로운 화자들을 식별하고, 신규 또는 임시 화자 프로파일들을 생성하고, 및/또는 분석 데이터베이스(104)의 화자 프로파일들을 업데이트한다. 오디오 데이터와 함께 수신된 가입자 식별자 또는 기타 메타데이터를 사용하여, 분석 서버(102)는 수신된 가입자 식별자와 관련된 성문(예를 들어, 의심되는 성문)을 식별하고, 임베딩 쌍 사이의 유사도 매트릭스를 생성하고, 임베딩 각각의 거리에 기초하여 유사한 임베딩을 클러스터링하고, 유사도 점수의 유사도 매트릭스를 생성하고, 다양한 임계값을 사용하여 최대 유사도 점수를 결정하고, 강한 임베딩 및 약한 임베딩을 결정하고, 약한 임베딩을 분석 데이터베이스(104)에 저장하고, 강한 임베딩을 사용하여 성문을 업데이트하고, 최대 유사도 점수와 연관된 성문을 평가함으로써 화자를 식별한다.

[0104] *임베딩 추출*

[0105] 분석 서버(102)는, 그 중에서도, 오디오 데이터 수집, 전처리 동작, 데이터 증대 동작, 임베딩 추출, 손실 함수 동작 및 분류 동작과 같은, 특정 동작들을 수행하도록 구성된 임의 개수의 계층들을 포함하는 하나 이상의 기계 학습 아키텍처들을 구현하기 위한 기계 실행 소프트웨어를 실행한다. 다양한 동작들을 수행하기 위해, 하나 이상의 기계 학습 아키텍처들은, 그 중에서도, 입력 계층들, 임베딩 추출기의 계층들, 완전-연결 계층들, 손실 계층들 및 분류기의 계층들과 같은, 임의 개수의 모델들 또는 계층들을 포함한다. 분석 서버(102)는 하나 이상의 기계 학습 모델들 및 계층들을 포함하는 오디오 처리 소프트웨어를 실행한다. 설명의 편의를 위해, 분석 서버(102)는 임베딩 추출기를 갖는 단일 기계 학습 아키텍처를 실행하는 것으로 설명되어 있지만, 일부 실시예들에서 다수의 기계 학습 아키텍처들(신경망 아키텍처들을 포함)이 사용될 수 있다.

[0106] 분석 서버(102)는 콘텐트 서버(111)로부터 오디오 신호를 수신하고 오디오 신호로부터 다양한 유형의 특징들을 추출한다. 분석 서버(102)는 오디오 신호의 백그라운드 노이즈, 정적, 및 화자들 사이를 구별하기 위해 오디오 이벤트 검출 또는 기타 음성 활동 검출을 수행한다. 예를 들어, 분석 서버(102)는 오디오 데이터를 전처리하거나(예를 들어, 오디오 신호를 필터링하여 노이즈를 감소시키고, 오디오 신호를 프레임 또는 서브프레임으로 파싱(parsing)하고, 다양한 정규화 또는 스케일링 동작들을 수행함), 음성 활동 검출(voice activity detection; VAD) 소프트웨어 또는 VAD 기계 학습을 실행하고, 및/또는 부분들(예를 들어, 프레임들, 세그먼트들)로부터 또는 실질적으로 모든 오디오 신호로부터 특징들(예를 들어, 하나 이상의 스펙트럼-시간적(spectro-temporal) 특징들)을 추출한다. 오디오 신호로부터 추출된 특징은 멜 주파수 켑스트럼 계수(Mel Frequency cepstrum coefficient; MFCC), 멜 필터 뱅크(Mel Filter bank), 선형 필터 뱅크(Linear filter bank), 병목 특징들 등을 포함할 수 있다.

[0107] 일부 실시예들에서, 콘텐트 서버(111)는 콘텐트 서버(111)가 최종 사용자 디바이스(114)로부터 수신하는 입력 오디오 신호에 VAD 또는 ASR 엔진을 적용한다. 콘텐트 서버(111)는 입력 오디오 신호의 음성 부분들을 분석 서버(102)로 전송하고, 분석 서버(102)는 입력 오디오 신호에 화자 인식(예를 들어, 임베딩 추출기)을 위한 기계 학습 모델을 적용한다.

[0108] 분석 서버(102)는 신경망 아키텍처(예를 들어, 심층 신경망(DNN), 컨볼루션 신경망(CNN)), 가우시안 혼합 모델들(GMM), 또는 기타 기계 학습 방법들을 사용하여 오디오 신호로부터 임베딩을 추출한다. 분석 서버(102)는 x-벡터, CNN 벡터, i-벡터 등을 사용하여 임베딩을 나타낼 수 있다.

[0109] 일 예로서, 분석 서버(102)는 오디오 신호로부터 음성 부분 세트 및 비음성 부분 세트를 파싱하는 VAD 동작들을 수행하기 위해 기계 학습 아키텍처를 훈련시킬 수 있다. 오디오 신호로부터 추출된 특징들에 VAD가 적용될 경우, VAD는 오디오 신호의 각 프레임(또는 서브 프레임)에 대한 이진 결과들(예를 들어, 음성 검출, 음성 검출 없음) 또는 지속적 값들(예를 들어, 음성 발생 확률들)을 출력할 수 있다. 오디오 신호의 음성 부분들은 발화라고 부를 수 있다. 오디오 신호는 다수의 화자들의 발화를 포함할 수 있다. 오디오 신호에는 중첩된 사운드(예를 들어, 발화 및 주변 백그라운드 노이즈)도 포함될 수 있다. 분석 서버(102)는 화자 검출 또는 다른 기존 화자 분할 솔루션들을 사용하여 발화의 시작과 끝을 결정할 수 있다.

[0110] 일부 실시예들에서, 콘텐트 서버(111)는 VAD 또는 ASR 기계 학습 모델을 실행하여 콘텐트 서버(111)가 최종 사용자 디바이스들(114)로부터 수신한 사용자 입력들의 음성 부분들을 식별한다. 이러한 실시예들에서, 콘텐트 서버(111)는 하나 이상의 화자들의 발화를 포함하는 오디오 신호들의 음성 부분들을 분석 서버(102)로 전송한다. 분석 서버(102)는 임베딩을 추출하기 전에 VAD 또는 ASR 동작들을 실행할 필요가 없다.

[0111] *재생 공격 대응*

[0112] 분석 서버(102)는 최종 사용자 디바이스(114)에 의해 캡처된 오디오 신호가 진정한 오디오 신호인지 또는 재생 공격(예를 들어, 물리적이고 반향적인 공간에서 마이크에 의해 캡처되고 재생 디바이스를 사용하는 최종 사용자 디바이스(114)의 마이크에 표현되는 오디오 신호)인지를 결정할 수 있다. 오디오 신호의 음성 부분들이 재생 공격인지 여부를 결정 시, 분석 서버(102)는 최종 사용자 디바이스(114) 유형 및 마이크 유형과 같은 화자-독립 특성들에 대해 콘텐트 시스템(110)을 수집, 결정 또는 쿼리할 수 있다. 추가로, 분석 서버(102)는 인바운드 오디오 신호의 다양한 아티팩트(artifact)들 및 부모의 성문에서의 대응하는 아티팩트 특징들에 기초하여 스푸핑(spoofing) 조건들을 식별하기 위해 인바운드 오디오 신호에 하나 이상의 훈련된 기계 학습 모델들을 적용할 수 있다.

[0113] *능동적 및 정적 등록*

[0114] 분석 서버(102)는 기계 학습 아키텍처를 사용하여 특정 등록자-화자에 대한 등록 단계 동안 특정 화자를 인식할 수 있다. 기계 학습 아키텍처는 등록자를 포함하는 음성 세그먼트들(또는 발화들)이 있는 등록자 오디오 신호들을 사용하여 등록자 음성 특징 벡터(때로는 "성문"이라고도 함)를 생성할 수 있다. 최종 사용자 디바이스들(114)과의 추후 능동적 또는 수동적 동작들 동안, 분석 서버(102)는 캡처된 오디오 신호들로부터 임베딩을 추출하고 추후 캡처된 오디오 신호들이 등록자를 포함하는지 여부를 확인하기 위해 임베딩을 성문과 비교한다.

[0115] 분석 서버(102)는 등록 단계(예를 들어, 미리 결정된 등록 시간) 동안 사용자들을 등록한다. 예를 들어, 분석 서버(102)는 새로운 최종 사용자 디바이스(114)의 초기화 동안 사용자들을 능동적으로 등록할 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 (예를 들어, 기존 성문을 업데이트하고, 새로운 사용자들에 대한 새로운 성문들을 생성하기 위해) 사용자들을 매년 능동적으로 등록할 수 있다.

[0116] 등록 단계 동안, 분석 서버(102)는 분석 서버(102)가 화자의 성문을 사용하여 특정 화자를 인식하기에 충분한 발화들(예를 들어, 오디오 신호의 음성 부분들)을 수신할 때까지 사용자들이 반복하는 등록 단계 동안 사용자에게 프롬프트할 수 있다.

[0117] 분석 서버(102)는 등록 단계 동안 등록 임베딩에 기초하여 성문을 생성한다. 등록 단계 동안, 분석 서버(102)는 분석 서버(102)가 특정 신호의 사용자를 수학적으로 식별할 수 있도록 하나 이상의 발화들로부터 임베딩들을 추출한다. 분석 서버(102)는 분석 서버(102)가 임계값을 초과하는 순(net) 음성의 지속기간을 수신할 때 등록 단계 동안 충분한 임베딩들이 추출되었다고 결정할 수 있다. 추가로 또는 대안으로, 분석 서버(102)가 분석 서버(102)가 미리 결정된 수의 등록 신호들(예를 들어, 2개의 지문 스캔 및 2개의 발화, 5개의 상이한 발화)을 가질 때 등록 단계 동안 충분한 임베딩이 추출되었다고 결정할 수 있다.

[0118] *지속적 및 수동적 등록*

[0119] 추가로 또는 대안으로, 분석 서버(102)는 지속적으로 사용자들을 등록한다. 정적 등록에서와 같이 미리 결정된(또는 지정된) 시간 기간에 사용자들을 등록하는 대신, 분석 서버(102)는 수신된 발화들과 관련된 성문들을 생성함으로써 언제든지 사용자들을 등록할 수 있다. 지속적 및 수동적 등록 동안 수신된 발화들이 임의의 지속기간 또는 품질의 발화들일 수 있으므로, 지속적 및 수동적 등록 동안 생성된 성문의 완성도는 다를 수 있다.

[0120] 예를 들어, 분석 서버(102)는 콘텐트 서버(111)로부터 포워딩된 오디오 신호(및 가입자 식별자)를 수신할 수 있다. 분석 서버(102)는 VAD 소프트웨어, 임베딩 추출기 모델, 또는 기타 기계 학습 모델들을 사용하여 오디오 신호로부터 특징들 및 임베딩들을 추출할 수 있다. 분석 서버(102)는 오디오 신호에 기계 학습 아키텍처를 적용하여 특정 화자에 대한 임베딩을 추출한다.

[0121] *비지도(Unsupervised) 클러스터링*

[0122] 일부 실시예들에서, 분석 서버(102)는 콘텐트 서버(111)로부터 포워딩된 가입자 식별자를 수신할 수 있다. 분석 서버(102)는 분석 데이터베이스(104)에 쿼리하고 수신된 가입자 식별자와 연관된 성문을 검색한다. 성문들은 가입자 식별자와의 연관성을 기반으로 하는 임베딩들에 대한 추정 성문들이다. 각 성문 및 관련 고유 화자 식별자는 하나 이상의 가입자 식별자에 링크된다.

[0123] 일부 실시예들에서, 분석 서버(102)는 콘텐트 서버(111)로부터 포워딩된 가입자 식별자를 수신하지 않을 수 있다. 추정 화자 클러스터들(화자 클러스터는 가입자 식별자와의 연관성에 기초한 추정 화자 클러스터임)과 연관된 임베딩들 및 성문들의 유사도를 평가하는 대신, 분석 서버(102)는 성문 세트로 임베딩들의 유사도를 평가할 것이다. 성문 세트는 분석 데이터베이스(104)로 최근에(예를 들어, 미리 결정된 양의 시간 내에) 전송된 성문들, 특정 화자 특성들과 연관된 성문들, 및 특정 화자-독립 특성들과 연관된 성문들을 포함할 수 있다.

[0124] 분석 서버(102)는 순차적 클러스터링 알고리즘들(예를 들어, k-평균 클러스터링)을 사용하여 클러스터링할(또는 아니면 유사한 임베딩들을 연관시킬) 수 있다. 분석 서버(102)는 유사도 매트릭스를 생성하고 다른 성문들에 대한 클러스터들(성문들) 각각의 유사도를 결정함으로써 성문들을 클러스터링할 수 있다. 일부 구성들에서, 분석 서버(102)는 성문 중심까지의 클러스터 내의 임베딩 각각의 거리를 평가함으로써 성문들에 대한 클러스터들의 유사도를 평가할 수 있다. 클러스터들의 유사도가 하나 이상의 임계값들을 충족하는 경우, 분석 서버(102)는 2개의 성문들을 단일 성문으로 병합할 수 있다(예를 들어, 성문의 평균을 취함).

[0125] 일부 구현들에서, 분석 서버(102)는 예를 들어 랜덤으로 중심을 생성하고 임베딩들을 중심과 연관시킴으로써 인바운드 및/또는 저장된 임베딩들을 클러스터링한다. 분석 서버(102)는 임베딩들과 중심 사이의 상대 거리들에 기초하여 임베딩들을 클러스터링한다. 분석 서버(102)는 중심과 연관된 임베딩 각각의 평균 거리를 최소화하는

것에 기초하여 중심을 새로운 상대 위치로 이동시킨다. 중심이 이동될 때마다, 분석 서버(102)는 임베딩들과 중심 사이의 거리들을 재계산한다. 분석 서버(102)는 중지 기준이 충족될 때까지 클러스터링 프로세스를 반복한다(예를 들어, 임베딩들이 클러스터들을 변경하지 않으며, 거리의 합이 최소화되고, 최대 반복 횟수에 도달함). 일부 구성들에서, 분석 서버(102)는 유클리드 거리를 사용하여 임베딩들과 중심들 사이의 거리들을 측정한다. 일부 구성들에서, 분석 서버(102)는 임베딩들의 특징들의 상관 관계에 기초하여 임베딩들과 중심들 사이의 거리들을 측정한다. 임베딩들과 중심들 사이의 거리는 유사도 점수를 사용하여 표시된다. 임베딩들이 중심과 유사할수록 유사도 점수가 높아진다. 분석 서버(102)는 유사도 매트릭스에서 각각의 임베딩과 중심들 사이의 유사도 점수들을 추적한다.

[0126] 추가로 또는 대안으로, 분석 서버(102)는 각 임베딩을 중심으로 처리할 수 있다. 분석 서버(102)는 다른 임베딩들까지의 중심 임베딩의 거리들에 기초하여 임베딩들을 클러스터링한다. 거리 측정들은 예를 들어, 다른 임베딩들까지의 최소 최대 거리, 다른 임베딩들까지의 최소 평균 거리, 및 다른 임베딩들까지의 최소 거리 제곱 합을 포함할 수 있다.

[0127] 분석 서버(102)는 또한 순차적 클러스터링 알고리즘들을 사용하여 성문들을 갖는 임베딩들을 클러스터링할 수 있다. 클러스터들은 성문이 화자 클러스터의 중심을 나타내는 특정 화자(예를 들어, 화자 클러스터)와 유사한 발화 모음을 나타낸다. 분석 서버(102)는 화자 식별자들을 사용하여 화자 클러스터들을 식별한다. 화자 식별자는 화자를 익명화하여, 하나의 화자 클러스터를 다른 화자 클러스터와 구별함으로써 화자의 개인 정보를 보호한다.

[0128] 일부 구성들에서, 분석 서버(102)는 메타데이터(또는 화자 특성들, 화자 식별자들)를 성문과 연관시킨다. 성문과 연관된 메타데이터는 오디오 신호의 품질을 포함할 수 있다. 예를 들어, 오디오 신호는 깨끗한 조건(예를 들어, 높은 신호 대 잡음비(signal-to-noise; SNR), 낮은 잔향 시간(reverberation time)(T60))의 발화들을 포함할 수 있다. 추가로 또는 대안으로, 오디오 신호는 노이즈가 많은 조건들(예를 들어, 낮은 SNR, 높은 T60)에서 발화들을 포함할 수 있다. 메타데이터는 또한 순 음성의 전체 지속시간을 포함할 수 있다. 분석 서버(102)는 화자 클러스터에서 발화 각각의 지속 시간들을 합산함으로써 음성의 지속시간을 결정할 수 있다. 메타데이터는 또한 화자 클러스터의 총 발화 수, 화자 특성들, 화자 식별자들 및/또는 화자-독립 특성들을 포함할 수 있다.

[0129] 분석 서버(102)는 또한 화자에 의해 입력되거나, 콘텐트 서버로부터 수신되거나, 다양한 기계 학습 모델들을 실행함으로써 식별된 정보에 기초하여 화자 특성들을 결정할 수 있다. 화자 특성들은, 그 중에서도, 예를 들어 화자의 연령, 화자의 성별, 화자의 감정 상태, 화자의 사투리, 화자의 억양, 화자의 발음을 포함할 수 있다. 일부 실시예들에서, 예를 들어, 분석 서버(102)는 Sadjadi 등의 "Speaker Age Estimation On Conversational Telephone Speech Using Senone Postterior Based I-Vectors" IEEE ICASSP, 2016, 및 Han 등의 “Age Estimation from Face Images: Human vs. Machine Performance,” ICB 2013에 의해 설명된 것들과 같은 기계 학습 모델들을 적용함으로써 특정 화자 연령 특성들을 식별할 수 있다. 그리고 일부 실시예들에서, 분석 서버(102)는 Buyukyilmaz 등의 "Voice Gender Recognition Using Deep Learning," Advances in Computer Science, 2016에 의해 설명한 것들과 같은 기계 학습 모델들을 적용함으로써 특정 화자 성별 특성들을 식별할 수 있다. 이 단락에서 위에서 언급한 참조 문헌 각각은 본원에 참조로 통합된다.

[0130] 분석 서버(102)는 하나 이상의 임베딩들에 기초하여 화자와 연관된 성문을 생성할 수 있다. 일부 구성들에서, 분석 서버(102)는 성숙 임베딩 클러스터(예를 들어, 등록 임베딩들에 기초한 클러스터)를 사용하여 성문을 생성한다. 성숙한 성문(또는 성숙한 임베딩 클러스터)은 분석 서버(102)가 화자 식별자를 사용하여 화자를 식별하기에 충분한 생체 인식 정보를 포함할 수 있다. 일부 구성들에서, 분석 서버(102)는 등록 임베딩들을 평균내어 성문을 생성할 수 있다. 일부 구성들에서, 분석 서버(102)는 메타데이터(또는 사용자 특성들, 화자 식별자들)를 성문과 연관시킬 수 있다.

[0131] 성문과 연관된 메타데이터는 성문에 있는 발화들(임베딩들로 표현됨)에 대한 정보를 포함할 수 있다. 메타데이터는 특정 임베딩(또는 성문)과 연관된 오디오 데이터의 품질을 포함한다. 예를 들어, 마이크는 깨끗한 조건들(예를 들어, 높은 신호 대 잡음비(SNR), 낮은 잔향 시간(T60))에서 오디오 데이터를 캡처할 수 있다. 마이크는 또한 노이즈가 많은 조건들(예를 들어, 낮은 SNR, 높은 T60)에서 오디오 데이터를 캡처할 수 있다. 메타데이터는 또한 순 음성의 전체 지속시간을 포함할 수 있다. 분석 서버(102)는 성문에서 발화 각각의 지속시간들을 합산함으로써 음성의 지속시간을 결정할 수 있다. 메타데이터는 클러스터의 총 발화 수를 포함할 수도 있다. 예를 들어, 분석 서버(102)는 하나의 10-초 발화를 수신할 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 5개의 1-초 발화를 수신할 수 있다. 메타데이터에는 사용자 특성들, 화자 식별자들 및/또는 화자-독립 특성들도 포함

될 수 있다.

[0132] 분석 서버(102)는 유사도 점수들의 유사도 매트릭스를 생성한다. 분석 서버(102)는 특정 화자에 대해 추출된 임베딩들과 데이터베이스에 저장된 다른 추정 화자들의 성문들 사이의 상대적 거리를 평가함으로써 각 특정 화자에 대한 유사도 점수들을 결정한다. 추가로 또는 대안으로, 분석 서버(102)는 코사인 유사도 또는 확률적 선형 판별 분석(probabilistic linear discriminant analysis; PLDA)을 사용하여 유사도 점수를 평가할 수 있다. 분석 서버(102)는 각각의 임베딩들과 화자 클러스터들 사이의 최대 유사도 점수를 식별함으로써 화자 클러스터들에 가장 유사한 임베딩들을 결정한다.

[0133] *강한 발화 및 약한 발화*

[0134] 분석 서버(102)는 다양한 임계값들을 사용하여 최대 유사도 점수를 평가할 수 있다. 예를 들어, 분석 서버(102)는 최대 유사도 점수를 더 낮은 임계값 및 더 높은 임계값 모두와 비교할 수 있다. 추가로 또는 대안으로, 분석 서버(202)는 하나 이상의 알고리즘들을 사용하여 더 높은 임계값과 더 낮은 임계값을 결합하여 특정 성문에 대한 최적의 임계값을 추정할 수 있다.

[0135] 분석 서버(102)가 특정 임베딩에 대한 최대 유사도 점수가 낮은 유사도 임계값에 미치지 못하다고 결정하는 경우, 분석 서버(102)는 화자가 새로운 알 수 없는 사용자일 가능성이 있다고 결정한다. 분석 서버(102)는 특정 임베딩을 사용하여 새로운 화자 프로파일을 생성할 것이다. 화자 프로파일에는 성문, 화자 클러스터(예를 들어, 성문과 연관된 임베딩들), 화자 식별자 및 메타데이터가 포함된다.

[0136] 분석 서버(102)가 최대 유사도 점수가 낮은 유사도 임계값을 충족한다고 결정하는 경우, 분석 서버(102)는 화자를 식별(또는 인증)할 수 있다. 추가로, 분석 서버(102)는 유사도 점수에 포함된 임베딩이 약한 발화에 기초하여 약한 임베딩이라고 결정한다. 약한 임베딩은 특정 화자 클러스터의 일부로서 약한 임베딩을 즉시 특성화하기 위한 해당 성문과 충분한 유사도가 없다. 분석 서버(102)는 분석 데이터베이스(104)에 약한 임베딩 세트를 저장 및/또는 업데이트할 수 있다.

[0137] 분석 서버(102)는 트리거링 기준에 기초하여 약한 임베딩 세트를 평가할 수 있다. 트리거링 기준의 비제한적인 예들에는 주기적인 약한 임베딩 평가 및 저장된 약한 임베딩들의 임계 수가 포함된다. 분석 서버(102)가 트리거링 조건을 식별하는 것에 응답하여, 분석 서버(102)는 약한 임베딩 세트와 연관된 성문에 대한 약한 임베딩 세트의 임베딩 각각의 유사도 점수를 재계산할 수 있다. 분석 서버(102)는 유사도 점수가 임계값을 초과하는 것에 응답하여 약한 임베딩들로 성문을 업데이트할 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 약한 임베딩 세트로부터 하나 이상의 약한 임베딩들을 제거할 수 있다.

[0138] 분석 서버(102)가 최대 유사도 점수가 낮은 유사도 임계값에 더하여 높은 유사도 임계값을 충족한다고 결정하는 경우, 분석 서버는 유사도 점수에 포함된 임베딩이 약한 임베딩 대신 강한 임베딩이라고 결정할 수 있다. 강한 임베딩은 성문과 매우 유사한(예를 들어, 상대적 거리와 관련하여 가까운) 임베딩이다. 분석 서버(102)는 임베딩을 포함하도록 화자 클러스터를 업데이트한다. 분석 서버(102)는 새로운 임베딩에 기초하여 성문을 재계산할 수 있다. 분석 서버(102)는 강한 임베딩들로 식별된 임베딩들과 강한 임베딩들로 식별되지 않은 임베딩들을 다르게 가중시킬 수 있다. 예를 들어, 분석 서버(102)는 임베딩들의 가중 평균을 취함으로써 성문을 업데이트할 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 알려진 화자와 관련된 강한 임베딩 목록을 업데이트할 수 있다.

[0139] 분석 서버(102)는 트리거링 조건들에 기초하여 약한 임베딩 세트에 대해 분석 데이터베이스(104)에 쿼리할 수 있다. 예를 들어, 분석 서버(102)는 주기적으로(예를 들어, 매주) 또는 약한 임베딩 세트가 미리 결정된 수에 도달할 때 분석 데이터베이스(104)에 쿼리할 수 있다. 분석 서버(102)는 약한 임베딩 세트와 연관된 성문에 대한 약한 임베딩 세트의 임베딩 각각의 유사도 점수를 재계산할 수 있다. 다양한 임계값들(예를 들어, 더 낮은 임계값 및/또는 더 높은 임계값)을 초과하는 최대 유사도 점수에 기초하여, 분석 서버(102)는 약한 임베딩으로 성문을 업데이트할 수 있다. 약한 임베딩은 시간이 지남에 따라 성문이 진화함(예를 들어, 노화되고, 더 정확해지고 성숙해짐)에 따라 강한 임베딩이 될 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 성문과 연관된 임베딩 세트에서 하나 이상의 약한 임베딩들을 제거할 수 있다.

[0140] *조건 종속적 적응 임계화*

[0141] 분석 서버(102)는 강한/약한 임베딩들을 적응적으로 결정하는데 사용되는 더 높은 임계값 및 더 낮은 임계값을 결정할 수 있다. 분석 서버(102)는 조건-종속적 적응 임계화를 사용하여 최대 유사도 점수와 관련된 성문 성숙

도에 기초하여 더 높은 임계값 및 더 낮은 임계값을 결정한다.

[0142] 분석 서버(102)가 성문이 성숙하다고 결정(예를 들어, 성숙도 임계값을 충족)하는 경우, 분석 서버(102)는 최대 유사도 점수를 평가할 때 높은 유사도 임계값 및 낮은 유사도 임계값을 사용하지 않을 수 있다. 예를 들어, 분석 서버(202)는 하나 이상의 알고리즘을 사용하여 높은 유사도 임계값들과 낮은 유사도 임계값들을 결합하여 특정 성문에 대한 최적의 임계값을 추정할 수 있다. 분석 서버(102)는 임베딩들 및 성문들의 최대 유사도 점수를 평가 시 최적의 임계값을 사용한다.

[0143] 분석 서버(102)는 하나 이상의 성숙도 인자들을 사용하여 성문이 성숙되어 있는지 여부를 결정할 수 있다. 성숙도 인자의 비제한적인 예들에는 등록 발화들의 수, 발화 전반에 걸친 순 음성의 전체 지속시간, 및 성문과 관련된 오디오 신호들로부터의 오디오 품질이 포함된다. 분석 서버(102)는 성문이 성숙한지 여부를 결정하기 위해 임의 개수의 알고리즘들을 사용할 수 있다. 예를 들어, 서버는 성숙도 인자(예를 들어, 발화들의 수)를 성숙도 인자(예를 들어, 발화들의 임계 수)에 해당하는 미리 구성된 성숙도 임계값과 비교한다. 다른 예로서, 서버는 성숙도 인자들을 통계적으로 또는 알고리즘적으로 결합하고 결합된 성숙도 인자를 결합된 성숙도 인자에 대응하는 미리 구성된 성숙도 임계값과 비교한다.

[0144] 분석 서버(102)가 성문이 성숙하지 않다(예를 들어, 성숙도 임계값을 충족하지 않다)고 결정하는 경우, 분석 서버(102)는 조건-종속적 적응 임계화를 사용하여 특정 성문들에 대한 높은 유사도 임계값들 및 낮은 유사도 임계값들을 결정한다.

[0145] 분석 서버(102)는 조건-종속적 적응 임계값을 활용하여 오수락률(a false acceptance rate; FAR)을 최소화한다. 오수락률은 분석 서버(102)가 화자를 잘못 인증 및/또는 식별하는 비율이다. (예를 들어, 관리 디바이스(103)를 사용하는) 관리자는 FAR(예를 들어, 0.5%, 1%, 2%, 3%, 4% 또는 5%)을 결정(또는 사전 구성)할 수 있다. 추가로 또는 대안으로, 콘텐트 시스템(110)이 화자 식별/인증과 관련된 특정 FAR들을 요청할 수 있거나, 또는 기계 학습 모델이 FAR을 알고리즘적으로 결정할 수 있다.

[0146] FAR은 성문의 성숙도에 기초하여 달라지기 때문에, 분석 서버(102)는 성문과 임베딩의 유사도를 평가할 때 서로 다른 임계값들(예를 들어, 높은 유사도 임계값들 및 낮은 유사도 임계값들)을 서로 다른 성문들에 적용한다. 일부 구성들에서, 분석 데이터베이스(104)는 하나 이상의 특정 조건들에 대한 다양한 FAR들에서 임계값들(예를 들어, 임계값 스케줄러)의 테이블을 저장할 수 있다.

[0147] *출력*

[0148] 분석 서버(102)는 오디오 신호에 기초하여 하나 이상의 화자 식별자들을 콘텐트 서버(111)에 전송할 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 화자 식별자들과 연관된 유사도 점수들을 콘텐트 서버(111)에 전송할 수 있다. 콘텐트 서버(111)는 수신된 화자 식별자를 사람의 화자(및 화자 식별자가 이전에 가입자 식별자와 연관되지 않은 경우, 가입자 식별자)와 매핑할 수 있다. 예를 들어, 분석 서버는 룩업 테이블을 사용하여 화자 식별자를 특정 사람 화자에 매핑할 수 있다. 분석 서버(102)는 또한 화자 식별자와 연관된 화자 프로파일(화자-독립 특성들, 화자 특성들, 메타데이터를 포함)을 전송할 수 있다.

[0149] 콘텐트 시스템(111)은 화자 식별자와 관련된 사용자의 선호도를 저장할 수 있다. 선호도의 비제한적인 예들에는 사용자가 역사적으로 보거나, 좋아하거나, 북마크한 콘텐트, 또는 아니면 사용자가 관심을 표시한 콘텐트 등이 포함될 수 있다. 콘텐트 시스템(111)은 수신된 화자 식별자에 기초하여 개인화된 콘텐트를 화자에게 스트리밍할 수 있다. 화자 식별자와 연관된 선호도가 저장되지 않은 경우(예를 들어, 새로운 사용자), 콘텐트 시스템(110)은 일반 콘텐트를 사용자에게 스트리밍할 수 있다.

[0150] 분석 서버(102)는 화자 또는 화자들의 현재 상황들과 환경을 나타내는 환경 설정을 식별할 수 있다. 분석 서버(102)에 의해 실행되는 기계 학습 모델은 예컨대 백그라운드 노이즈 또는 분류할 수 있는 특정 사운드(예를 들어, 식기세척기, 트럭들)와 같은, 오디오 이벤트 분류 모델 및/또는 환경 분류 모델을 포함하거나 또는 인바운드 신호에 대한 압도적인 양의 에너지를 포함할 수 있다. 분석 서버(102)는 특정 콘텐트 특성들과 연관된 환경 설정의 표시자와 함께 화자 식별자들을 콘텐트 서버(111)에 전송할 수 있다. 예를 들어, 성인 화자들만 있는 레스토랑이나 파티에서 스마트 TV(114a)와 상호작용하는 화자는 화자가 아동 화자들을 포함하는 거실의 스마트 TV(114a)와 상호작용하는 다른 상황들로부터 콘텐트 서버(111)가 최종 사용자 디바이스(114)에 서로 다른 제안된 콘텐트를 생성하게 할 수 있다.

[0151] 일부 구성들에서, 콘텐트 서버(111)는 분석 서버(102)의 출력을 참조하여 특정 가입 및 인증된 사용자들의 수에 대한 액세스를 제한한다. 분석 서버(102)가 콘텐트 서버(111)로부터 가입자 식별자를 수신했다면, 분석 서버

(102)는 화자 프로파일(및 연관된 화자 식별자, 화자 특성들, 화자-독립 특성들, 메타데이터)을 사용하여 화자가 특정 가입자 식별자와 관련된 인증 규칙들 및 제한 사항들에 기초하여 가입자 계정에 액세스하도록 인증되었는지 여부를 결정할 수 있다. 분석 서버(102)는 화자 식별자가 특정 가입자 식별자에 대해 인증되었는지 여부의 표시를 콘텐트 시스템(110)에 전송할 수 있다. 추가로 또는 대안으로, 콘텐트 서버(111)는 제한된 특정 콘텐트에 대한 화자 프로파일 정보를 이용하여 화자를 인증할 수 있다. 예를 들어, 화자 프로필에 식별된 화자의 연령은 (예를 들어, 부모 통제 또는 특정 연령에 적합한 등급에 기초하여) 연령 제한 콘텐트를 소비하도록 한 명 이상의 화자들을 인증할 수 있다.

[0152] 일부 실시예들에서, 분석 서버(102)는 연령 제한 콘텐트를 활성화하거나 콘텐트 서버(111)에게 분석 서버(102)가 콘텐트 서버(111)로부터 수신한 입력 데이터(예를 들어, 인바운드 오디오 신호, 인증 데이터, 메타데이터, 최종 사용자 디바이스 데이터(114))에 기초하여 연령 제한 콘텐트를 활성화하도록 지시한다.

[0153] *라벨 정정*

[0154] 일부 구성들에서, 분석 시스템(101)은 라벨 식별자들(예를 들어, 화자 식별자들, 가입자 식별자들)을 정정한다. 분석 시스템(101)은 성문들을 클러스터링(예를 들어, 계층적 클러스터링)함으로써 라벨들을 정정한다. 라벨 식별자들을 정정하면 작은 누적 식별/인증 오류들이 발생할 가능성을 최소화하여, 화자 클러스터들의 순도를 높인다.

[0155] 분석 서버(102)는 배치 단계 동안 생성된 성문들의 라벨 식별자들을 정정할 수 있다. 분석 서버(102)는 라벨 식별자들이 콘텐트 서버(111)로 전송되기 전에 라벨 식별자들을 정정한다. 분석 서버(102)는 또한 데이터베이스(예를 들어, 콘텐트 데이터베이스(112), 분석 데이터베이스(104))에 저장된 성문들의 라벨 식별자들을 정정할 수 있다.

[0156] 분석 서버(102)는 라벨 정정을 트리거하는 식별 기준에 응답하여 라벨 식별자들을 정정한다. 트리거링 기준의 비제한적인 예들에는 주기적인 시간 간격 또는 미리 구성된 라벨-정정 스케줄; 클러스터링 또는 재-클러스터링 동작 실행; 특정 개수의 새로운 화자 식별자들의 식별 아니면 수신; 또는 특정 개수의 성문들의 생성이 포함된다.

[0157] 일부 구성들에서, 분석 서버(102)가 트리거링 기준을 식별하더라도, 분석 서버(102)는 라벨 식별자들을 정정하지 않도록 결정할 수 있다. 분석 서버(102)는 분석 서버(102)에 의해 검색된 성문들이 높은 신뢰도 값과 연관되는 경우 라벨 식별자들을 정정하지 않기로 결정할 수 있다. 분석 서버(102)는 분석 서버(102)가 능동적 등록 단계 또는 수동적 등록 단계 동안 성문을 생성했는지 여부에 기초하여 신뢰도 값을 연관시킬 수 있다. 분석 서버는 능동적 등록 단계 동안 생성된 성문들이 수동적 등록 단계 동안 생성된 성문들보다 신뢰도 값이 더 높다고 결정할 수 있다. 능동적 등록 임베딩들에 기초하여 생성된 성문은 순수하고 성숙한 것으로 간주될 수 있다. 추가로 또는 대안으로, 분석 서버(102)는 분석 시스템(101)이 느리게 실행되는 경우 라벨 식별자들을 정정하지 않도록 결정할 수 있다.

[0158] 분석 서버(102)는 데이터베이스(예를 들어, 콘텐트 데이터베이스(112) 또는 분석 데이터베이스(104))를 쿼리하고 화자 프로파일 세트(성문들, 임베딩들을 포함하는 화자 클러스터들, 가입자 식별자들 또는 화자 식별자들을 포함)을 검색할 수 있다. 화자 프로파일 세트는 최근에 액세스 및/또는 수정된 화자 프로파일들(예를 들어, 지난 2일 동안 분석 서버에 의해 검색된 화자 프로파일들), 특정 화자 특성들과 관련된 화자 프로파일들, 특정 화자 독립 특성들과 관련된 화자 프로파일들 및/또는 다른 메타데이터와 관련된 화자 프로파일들일 수 있다.

[0159] 일부 구성들에서, 분석 서버(102)는 오디오 신호로부터 추출된 임베딩들과 검색된 성문들 사이의 쌍별 유사도를 계산할 수 있다. 검색된 성문들은 오래된 라벨링된 세트로 간주되고, 오디오 신호로부터의 임베딩들은 새로운 익명 세트로 간주된다. 추가로 또는 대안으로, 분석 서버(102)는 검색된 성문들 사이의 쌍별 유사도를 계산할 수 있다.

[0160] 분석 서버(102)는 새로운 익명 세트의 성문들과 오래된 라벨링된 세트의 성문들의 유사도에 기초하여 오래된 라벨링된 세트와 연관된 라벨 식별자들을 새로운 익명 세트로 마이그레이션(migrate)한다. 분석 서버(102)는 (예를 들어, 유클리드 거리 측정, 상관 관계 기반 측정에 따라) 가까운 성문들을 평가함으로써 성문들의 유사도를 결정한다. 분석 서버(102)가 성문들이 가깝다고 결정하는 경우(예를 들어, 상대적 거리가 임계값을 충족함), 성문들 및 연관된 라벨 식별자들이 병합될 수 있다. 분석 서버(102)는 새로운 익명 세트의 화자 식별자 및/또는 가입자 식별자가 오래된 라벨링된 세트의 화자 식별자 및/또는 가입자 식별자로 대체되도록 오래된 라벨링된 세트와 연관된 라벨 식별자들을 새로운 익명 세트로 마이그레이션할 수 있다. 라벨 식별자들을 마이그레이션하는

것 외에, 분석 서버(102)는 오래된 라벨링된 세트의 중심과 새로운 익명 세트의 중심을 평균냄으로써 병합된 성문들에 대한 새로운 중심을 결정할 수 있다. 일부 구성들에서, 분석 서버(102)는 오래된 라벨링된 세트의 라벨들을 새로운 익명 세트로 마이그레이션하기 전에 사용자 특성들을 비교할 수 있다. 일부 구성들에서, 분석 서버(102)는 마이그레이션된 라벨들로 분석 데이터베이스 및/또는 콘텐트 데이터베이스를 업데이트한다.

[0161] *사용자 및 부모 통제 인증*

[0162] 본원에 논의된 바와 같이, 분석 서버(102)는 추출된 임베딩의 유사도를 분석 데이터베이스(104)에 저장된 임베딩들/성문들과 비교함으로써 최종 사용자 디바이스(114)와 상호작용하는 사용자의 ID를 결정할 수 있다. 분석 서버(102)는, 하나 이상의 사용자들을 식별 시, 사용자 식별자, 사용자 특성들(예를 들어, 연령, 성별, 감정, 사투리, 억양 등), 사용자-독립 특성들, 및/또는 메타데이터를 콘텐트 서버(111)에 전송할 수 있다. 일부 구성들에서, 분석 서버(102)(또는 분석 서버(102)로부터 전송된 정보를 사용하는 콘텐트 서버(111))는 전송된 정보를 사용하여 식별된 사용자들을 인증할 수 있다. 예를 들어, 사용자 연령은 특정 연령 제한을 초과하는 콘텐트를 시청하는 사용자를 인증할 수 있다.

[0163] 일부 구성들에서, 분석 서버(102)(또는 콘텐트 서버(111))는 식별된 사용자들에 기초하여 사용자가 콘텐트를 시청하도록 인증되었는지 여부를 결정할 수 있다. 예를 들어, 분석 서버(102)는 텔레비전을 시청하고 있는 8세 소년을 식별할 수 있다. 분석 서버(102)는 오디오 신호의 분석에 기초하여 상승된 권한을 갖는 다른 화자를 식별하고, 여기서 분석 서버(102)는 2명의 화자들에 대해 추출된 임베딩들과 일치하는 성문들을 갖는 2명의 화자들의 화자 프로파일들을 식별하였다. 예를 들어, 분석 서버(102)는 8세 소년과 동일한 오디오 신호에서 아동의 부모를 식별한다. 8세 소년과 매우 가까운 성인 남성의 존재는 분석 서버(102)(또는 콘텐트 서버(111))가 8세 소년에게 특정 콘텐트를 시청하도록 하는 권한을 부여하는 결과를 초래한다.

[0164] **분석 데이터베이스**

[0165] 분석 데이터베이스(104)는 특정 콘텐트 시스템(110), 가입자 식별자들(예를 들어, 룩업 테이블)과 연관된 화자 식별자들(및 성문들), 추출된 임베딩들(예를 들어, 약한 임베딩들), 사용자 특성들, (예를 들어, VAD 동작을 실행하기 위한, 임베딩을 추출하기 위한) 훈련된 기계 학습 모델들 등에 대한 FAR들을 저장할 수 있다.

[0166] 분석 데이터베이스(104)는 클러스터의 임베딩들이 하나 이상의 임계값들을 충족하는 경우(예를 들어, 분석 서버(102)가 클러스터가 성숙한 등록 클러스터라고 결정하는 경우) 성문으로서 클러스터링된 임베딩들을 저장할 수 있다. 분석 서버(102)는 클러스터의 발화 지속시간(클러스터의 임베딩들로 표현됨)이 임계값을 충족하거나, 클러스터의 발화들의 수가 임계값을 충족하거나, 일부 조합 등을 충족하는 경우 등록 클러스터가 성숙하다고 결정할 수 있다.

[0167] 추가로 또는 대안으로, 분석 데이터베이스(104)는 클러스터링된 임베딩이 화자 식별자 및/또는 가입자 식별자와 연관되지 않은 경우 클러스터링된 임베딩을 성문으로 저장할 수 있다. 분석 데이터베이스(104)는 성문이 성숙되지 않은 경우에도 성문을 저장할 수 있다.

[0168] 일부 구성들에서, 분석 데이터베이스(104)는 저장된 성문들을 퍼징(제거, 삭제)할 수 있다. 예를 들어, 분석 데이터베이스(104)는 가입자 식별자와 관련된 화자 식별자들을 제거하라는 명령어들을 콘텐트 서버(111)로부터 수신할 수 있다(예를 들어, 가입자는 콘텐트 시스템(110)의 서비스들을 구동 취소하기로 결정할 수 있음). 추가로 또는 대안으로, 분석 데이터베이스(104)는 미리 결정된 양의 시간이 주어지면 저장된 성문을 제거할 수 있다. 추가로 또는 대안으로, 분석 데이터베이스(104)는 분석 데이터베이스(104)(또는 분석 서버(102))가 성문이 특정 기준을 충족한다고 결정하면 저장된 성문을 제거할 수 있다. 예를 들어, 성문 클러스터의 임베딩들은 합성 음성에 기초한다.

[0169] 동작 시, 분석 데이터베이스(104)는 콘텐트 서버(111)로부터 오디오 데이터 및 가입자 식별자를 수신할 수 있다. 분석 데이터베이스(104)는 가입자 식별자와 관련된 화자 식별자들(및 관련 성문들)을 조회할 수 있다. 분석 데이터베이스(104)는 또한 성문들 및 임의의 사용자 특성들과 연관된 더 높은/낮은 임계값들을 조회할 수 있다. 분석 데이터베이스(104)는 또한 콘텐트 시스템(110)과 연관된 FAR들을 조회할 수 있다. 분석 데이터베이스(104)는 룩업 테이블에서 성문 엔트리들, 더 높은/낮은 임계값들, FAR들, 및 사용자 특성들을 업데이트할 수 있다. 분석 데이터베이스(104)는 룩업 테이블에 성문 엔트리들, 가입자 식별자들, 더 높은/낮은 임계값들, FAR들, 및 사용자 특성들을 추가할 수 있다.

[0170] **관리 디바이스**

[0171] 분석 시스템(101)의 관리 디바이스(103)는 분석 시스템(101)의 직원이 다양한 관리 태스크들 또는 사용자 실행 식별, 보안 또는 인증 동작들을 수행할 수 있도록 하는 컴퓨팅 디바이스이다. 관리 디바이스(103)는 프로세서 및 소프트웨어를 포함하고, 본원에 설명된 다양한 태스크들 및 프로세스들을 수행할 수 있는 임의의 컴퓨팅 디바이스일 수 있다. 관리 디바이스(103)의 비제한적인 예들에는 서버, 개인용 컴퓨터, 랩탑 컴퓨터, 태블릿 컴퓨터 등을 포함할 수 있다. 동작 시, 사용자는 분석 서버(102)와 같은 시스템(100)의 다양한 컴포넌트들의 동작들을 구성하기 위해 관리 디바이스(103)를 사용하고, 또한 사용자들이 시스템(100)의 다양한 컴포넌트들에 쿼리들 및 명령어들을 발행하도록 허용할 수 있다. 예를 들어, 관리 디바이스(103)는 콘텐트 시스템(110)과 연관된 FAR을 결정하는 데 사용될 수 있다.

[0172] **미디어 콘텐트 시스템을 위한 클라이언트측 오디오 처리**

[0173] 설명 및 이해의 용이함을 위해, 본원에 설명된 실시예들은 부분적으로 화자 발화 및 음성 입력들에 따라 동작하는 콘텐트 전달 시스템들의 맥락에서 이러한 기술을 사용하는 것을 언급한다. 그러나 실시예들은 이에 제한되는 것은 아니며 수동적(또는 능동적) 등록, 지속적(또는 정적) 등록, 또는 다중 화자 음성 생체 인식의 지속적 식별/인증으로부터 이익을 얻을 수 있는 임의 개수의 시스템들 또는 제품들에 사용될 수 있다. 예를 들어, 본원에 설명된 다중 화자 음성 생체 인식 시스템들 및 동작들의 식별/인증은 오디오 입력들(예를 들어, 자동차들 또는 스마트 기기들 또는 콜 센터들과 같은 에지-디바이스들/IoT 디바이스들)을 수신하고 식별하는 모든 시스템에서 구현될 수 있다.

[0174] 게다가, 본원의 실시예들은 화자들을 특정 알려진 또는 알 수 없는 사용자들로 식별하기 위해 오디오 처리 동작들을 사용한다. 그러나 실시예들은 음성 생체 인식으로만 제한되는 것은 아니며, 화자들을 특정 사용자들로 식별하기 위해 임의 개수의 추가 유형의 생체 인식들을 수집 및 처리할 수 있다. 실시예들이 수집 및 처리할 수 있는 추가 유형의 생체 인식들의 비제한적인 예들에는 모니터링되는 시스템(예를 들어, 키 누르기, 메뉴 액세스, 콘텐트 선택, 입력 또는 선택 속도) 또는 임의의 생체 인식 정보의 조합에 액세스할 경우의 눈 스캔(예를 들어, 망막 또는 홍채 인식), 얼굴(예를 들어, 얼굴 인식), 지문 또는 손자국(예를 들어, 지문 인식), 사용자 행동(예를 들어, "행동 인쇄")이 포함된다.

[0175] 도 2는 기계 학습 모델들 및 기타 기계 학습 아키텍처가 로컬 디바이스 상에 구현되는, 오디오 처리 기계 학습 동작들을 사용하는 시스템(200)의 컴포넌트들을 도시한다. 시스템(200)은 분석 시스템(201), 미디어 콘텐트 시스템(210), 및 최종 사용자 디바이스들(214)을 포함한다. 분석 시스템(201)은 분석 서버(202), 분석 데이터베이스(204), 및 관리 디바이스(203)를 포함한다. 콘텐트 시스템(210)은 콘텐트 서버들(211) 및 미디어 콘텐트 데이터베이스들(212)을 포함한다. 실시예들은 추가 또는 대안적인 컴포넌트들을 포함하거나 도 2의 컴포넌트들로부터 특정 컴포넌트들을 생략할 수 있으며, 여전히 본 개시의 범위에 속한다. 예를 들어, 다수의 콘텐트 시스템들(210)을 포함하거나 분석 시스템(201)이 다수의 분석 서버들(202)을 갖는 것이 일반적일 수 있다. 추가로 또는 대안으로, 분석 시스템(201), 또는 분석 시스템(201)의 일부가 최종 사용자 디바이스(214)에 임베딩될 수 있다. 실시예들은 본원에 설명된 다양한 특징들 및 태스크들을 수행할 수 있는 임의 개수의 디바이스들을 포함하거나 아니면 구현할 수 있다. 예를 들어, 도 2는 분석 데이터베이스(204)와 별개의 컴퓨팅 디바이스로서의 분석 서버(202)를 도시한다. 일부 실시예들에서, 분석 데이터베이스(204)는 분석 서버(202)와 통합된다.

[0176] 도 2의 시스템(200)은 도 1의 시스템(100)과 동작이 유사하나, 도 2에서 최종 사용자 디바이스(214)는 다양한 오디오 처리 및 데이터 분석 동작들을 수행한다. 예를 들어, 최종 사용자 디바이스들(214)과 사용자의 능동적인 상호작용에 응답하여, 최종 사용자 디바이스들(214)은 사용자 입력 데이터(예를 들어, 오디오 신호)를 캡처할 것이다. 캡처된 오디오 신호를 콘텐트 시스템(210)으로 포워딩하는 최종 사용자 디바이스들(214) 대신(도 1에서와 같이), 오디오 신호를 콘텐트 시스템(210) 및 분석 시스템(201) 양쪽(또는 어느 한쪽)에 포워딩하기 전에 하나 이상의 기계 학습 아키텍처들이 오디오 신호에 적용된다.

[0177] 일부 구성들에서, 최종 사용자 디바이스들(214)은 다양한 형태의 메타데이터를 식별하거나 추출할 수 있고; 오디오 신호를 수정, 변환 또는 강화하고; 오디오 신호에서 음성을 식별하기 위한 소프트웨어 프로세스들 및 서비스들을 호스팅 및 실행하고; 오디오 신호에서 화자들과 관련된 생체 인식 특징들을 추출하고; 오디오 신호에서 화자들을 식별/인증한다. 추가로 또는 대안으로, 최종 사용자 디바이스들(214)은 오디오 신호를 필터링(오디오 신호의 노이즈 제거)하고, 오디오 신호의 형식을 변환하고, 오디오 신호를 파싱(또는 분할)하고, VAD 소프트웨어(또는 VAD 기계 학습)를 실행하고, ASR을 수행하고, 오디오 신호를 스케일링할 수 있다.

[0178] 일부 구성들에서, 최종 사용자 디바이스들(214)은 미리 결정된 양의 시간 동안 버퍼 또는 캐시 메모리와 같은, 데이터베이스 또는 비일시적 기계 판독 가능 저장 매체에 오디오 신호를 저장한다. 트리거링 조건에 응답하여,

최종 사용자 디바이스(214)는 저장된 오디오 신호를 검색하고 저장된 오디오 신호, 및 트리거링 조건과 연관된 오디오 신호를 분석 시스템(201)으로 포워딩할 수 있다. 예를 들어, 사용자가 사운드를 능동적으로 캡처하기 위해 최종 사용자 디바이스(214)를 동작시킬 경우, 사용자 최종 디바이스(214)는 사용자가 최종 사용자 디바이스(214)를 능동적으로 동작시키기 직전에 최종 사용자 디바이스(114)에 의해 캡처된 저장된 오디오 신호들을 검색한다. 일부 실시예들에서, 최종 사용자 디바이스들(214)은 오디오 신호들을 처리하고 오디오 신호들에서 화자들을 식별하기 위해 오디오 신호들을 분석 시스템(201)에 포워딩한다. 추가로 또는 대안으로, 최종 사용자 디바이스(214)는 오디오 신호들을 처리하고 오디오 신호들에서 화자들을 식별할 수 있다. 최종 사용자 디바이스들(214)은 오디오 신호들로부터 임베딩들을 추출함으로써 오디오 신호들을 처리할 수 있다.

[0179] 동작 시, 최종 사용자 디바이스들(214)은 오디오 신호들에 다양한 기계 학습 동작들을 적용할 수 있다. 예를 들어, 최종 사용자 디바이스들(214)은 오디오 신호로부터 특징들을 추출하도록 구성된 VAD 소프트웨어 또는 기타 기계 학습 아키텍처들을 실행함으로써 임베딩들을 추출할 수 있다. 일부 실시예들에서, 최종 사용자 디바이스들(214)은 분석 시스템(201)이 도 1에 설명된 바와 같이 분석 데이터베이스(204)로부터 검색된 성문들로 임베딩들을 클러스터링할 수 있도록 추출된 임베딩들을 분석 시스템(201)에 포워딩할 수 있다. 분석 서버(202)는 저장된 성문들에 대한 오디오 신호의 발화의 유사도에 기초하여 화자들을 식별할 수 있다. 분석 서버(202)는 화자 식별자 및 기타 메타데이터를 콘텐트 시스템(210)에 포워딩할 수 있다.

[0180] 추가로 또는 대안으로, 성문들이 최종 사용자 디바이스(214)에 저장되면, 최종 사용자 디바이스(214)는 저장된 성문과 임베딩들을 클러스터링하고 최종 사용자 디바이스들(214)에 저장된 하나 이상의 성문들과 비교하여 각 추출된 임베딩에 대한 유사도 점수들을 설명하는 유사도 매트릭스를 생성할 수 있다. 추가로 또는 대안으로, 최종 사용자 디바이스들(214)은 분석 시스템(201)에서 분석 데이터베이스(204)에 쿼리하고 도 1에 설명된 바와 같이 가입자 식별자 또는 화자 특성들과의 연관성에 기초하여 추출된 임베딩들에 대한 의심스러운 성문들 또는 추정 성문들을 검색할 수 있다.

[0181] 일부 구성들에서, 최종 사용자 디바이스들(214)은 분석 시스템(201)이 도 1에 설명된 바와 같이 조건-종속적 적응 임계화를 사용하여 최대 유사도 점수를 높은 유사도 임계값들 및 낮은 유사도 임계값들과 비교할 수 있도록 유사도 매트릭스들을 분석 시스템(201)에 포워딩한다. 분석 서버(211)는 약한 발화들, 강한 발화들을 결정하고 성문들을 업데이트할 수 있다. 추가로 또는 대안으로, 최종 사용자 디바이스(214)는 도 1에 설명된 바와 같이 분석 시스템(201)에서 분석 데이터베이스(204)에 쿼리하고, FAR들, 성숙도 임계값들 및 유사도 임계값들을 검색할 수 있다. 최종 사용자 디바이스들(214)은 유사도 매트릭스의 최대 유사도 점수들을 유사도 임계값들과 비교하는 것에 기초하여 성문들을 업데이트할 수 있다. 최종 사용자 디바이스들(214)은 업데이트된 성문들을 저장할 수 있고 또한 업데이트된 성문들을 분석 데이터베이스(204)로 전송할 수 있다.

[0182] 일부 구성들에서, 도 1에 설명된 바와 같이, 트리거링 기준에 응답하여, 분석 서버(202)는 라벨 정정을 실행하여 라벨 식별자들을 정정할 수 있다. 추가로 또는 대안으로, 최종 사용자 디바이스들(214)은 라벨 정정들을 실행하고 업데이트된 라벨 식별자들을 분석 시스템(201) 및/또는 콘텐트 시스템(210)에 포워딩할 수 있다.

[0183] 최종 사용자 디바이스들(214)이 화자 식별자를 사용하여 오디오 신호로부터 화자를 식별하는 경우, 최종 사용자 디바이스들(214)은 콘텐트 시스템(210)이 식별된 화자에게 스트리밍할 개인화된 콘텐트를 결정할 수 있도록 화자 식별자 및 임의의 메타데이터를 콘텐트 시스템(210)으로 전송할 수 있다. 최종 사용자 디바이스들(214)은 또한 화자 프로파일(화자 식별자, 추출된 임베딩들, 메타데이터, 성문들을 포함)을 분석 시스템(201)(예를 들어, 분석 데이터베이스(204))에 전송할 수 있다.

[0184] **예시적인 동작들**

[0185] **오디오 처리 인증 시스템에서의 능동적 및 정적 등록**

[0186] 도 3은 서버가 사용자를 식별(또는 인증)하기 위해 진행하는 단계들을 설명한다. 도 3a는 일 실시예에 따른 등록 단계 동안 사용자들을 능동적으로 등록하기 위한 방법(300a)의 동작 단계들을 도시한다. 도 3b는 배치 단계 동안 사용자를 식별(또는 인증)하기 위한 방법(300b)의 동작 단계들을 도시한다. 분석 시스템의 서버(예를 들어, 분석 서버)는 아래에 설명된 방법들(300a, 300b)을 수행하는 기계 판독 가능 소프트웨어 코드를 실행하지만, 임의 개수의 컴퓨팅 디바이스들의 하나 이상의 프로세서들이 방법들(300a, 300b)의 다양한 동작들을 수행할 수 있다. 일부 실시예들은 방법들(300a, 300b)에서 설명된 방법들보다 추가 동작, 적은 동작 또는 다른 동작을 포함할 수 있다.

[0187] 도 3a를 참조하면, 단계(302)에서, 서버는 사용자에게 등록 신호에 대해 프롬프트한다. 일부 구성들에서, 서버

는 등록 신호들을 사용자에게 한 번 또는 주기적 시간 간격으로(예를 들어, 6개월마다) 프롬프트할 수 있다. 일부 구성들에서, 서버는 예를 들어 디바이스가 처음으로 시작될 때, 미리 결정된 간격으로, 또는 사용자가 새로운 화자-사용자를 등록(register)/등록(enroll)하기 위해 구성 인터페이스에 액세스할 때 사용자에게 등록 신호들에 대해 프롬프트할 수 있다. 일부 구성들에서, 서버는 등록 단계를 실행하기 위한 (사용자 또는 기타 관리자로부터의) 명령어들에 기초하여 사용자에게 등록 신호들에 대해 프롬프트할 수 있다. 프롬프트들은 사용자가 지문 센서에 손가락을 대도록 프롬프트하고, 사용자가 특정 구문을 말하도록 프롬프트하고, 사용자가 자연스럽게 말하도록 프롬프트하고, 사용자가 디스플레이의 디지털 경계 상자의 경계들 내에 나타나도록 프롬프트하는 등을 포함할 수 있다.

[0188] 단계(304)에서, 서버는 등록 신호를 수신한다. 등록 신호는 지정된 등록 단계(예를 들어, 사용자의 능동적 등록) 동안 수신된 신호이다. 최종 사용자 디바이스 및 서버가 등록 단계 동안 그리고 (단계 302에서와 같이) 프롬프트에 응답하여 등록 신호들을 수신하기 때문에 등록 신호들은 다른 유형의 신호들과 구별될 수 있다. 예를 들어, 서버는 사용자가 최종 사용자 디바이스의 버튼을 누르고 최종 사용자 디바이스가 서버에 전송하는 오디오 또는 시각적 프롬프트에 응답하여 특정 발화를 말하는 것에 응답하여 등록 신호를 수신한다. 서버는 오디오 신호를 직접 수신하거나 중개 서버(intermediary server)(예를 들어, 콘텐트 서버)를 통해 수신한다. 일부 구성들에서, 서버는 여러 등록 신호들(예를 들어, 신호의 다수의 음성 부분들(발화들)) 및 등록 발화들 외에 생체 인식 데이터(예를 들어, 하나 이상의 지문 각도들)를 등록하기 위한 임의 개수의 생체 인식 데이터를 수신한다.

[0189] 일부 구성들에서, 서버는 기계 판독 가능 형식의 오디오 데이터를 포함하는 데이터 파일 또는 데이터 스트림인 오디오 신호를 수신한다. 오디오 신호는 임의 명수의 화자들에 대한 임의 개수의 화자 발화들을 포함하는 오디오 레코딩을 포함한다. 오디오 데이터에는 또한 오디오 신호와 함께 수신된 데이터 또는 메타데이터가 포함될 수 있다. 예를 들어, 오디오 데이터에는 화자 관련 정보(예를 들어, 사용자/화자 식별자, 가입자/가구 식별자, 사용자 생체 인식) 또는 통신 프로토콜들 또는 매체와 관련된 메타데이터(예를 들어, TCP/IP 헤더 데이터, 전화번호/ANI)가 포함될 수 있다.

[0190] 단계(306)에서, 서버는 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기)을 포함하는 기계 학습 아키텍처를 적용하여 등록 신호로부터 임베딩들을 추출한다. 임베딩들은 등록 신호에서 생체 인식 정보(또는 생체 인식 정보의 특징들)의 수학적 표현들이다. 서버는 멜 주파수 켑스트럼 계수들(MFCC들), 멜 필터 뱅크들, 선형 필터 뱅크들, 병목 특징들 등을 포함하는 발화들로부터 특징들을 추출할 수 있다. 서버는 특징들을 추출하고 스피커 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 입력 오디오 신호로부터 특징들을 추출한다.

[0191] 등록 신호의 유형은 서버가 임베딩들을 추출하는 방법을 지시할 수 있다. 예를 들어, 사용자는 지문을 등록 신호로 제공할 수 있다. 다른 예들에서, 서버는 능선, 골, 세부 사항을 포함하여 디지털 지문 이미지와 관련된 특징들을 추출할 수 있고; 및/또는 서버는 이미지 데이터의 특징들을 추출하도록 구성된 기계 학습 모델들을 사용하여 특징들을 추출할 수 있다. 서버는 화자 임베딩들과 함께, 사용자 인식에 사용되는 특정 유형의 생체 인식에 대해 특징들을 추출하고 해당 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 다양한 유형의 특징들을 추출한다.

[0192] 결정 단계(308)에서, 서버는 등록이 성숙되었는지 여부를 결정한다. 서버는 임베딩들 또는 기타 정보의 임계 수를 충족하기에 충분한 임베딩들(또는 생체 인식 정보)을 추출한 경우 등록이 성숙되었다고 결정할 수 있다. 예를 들어, 서버는 서버가 순 음성의 미리 결정된 지속시간을 수신할 때 등록이 성숙되었다고 결정할 수 있다. 추가로 또는 대안으로, 서버가 미리 결정된 수의 다양한 등록 신호들(예를 들어, 2개의 지문 스캔 및 2개의 발화, 5개의 상이한 발화)을 수신할 때 등록이 성숙되었다고 결정할 수 있다. 성숙한 성문(또는 성숙한 임베딩 클러스터)은 분석 서버(102)가 화자 식별자를 사용하여 화자를 식별하기에 충분한 생체 인식 정보를 포함할 수 있다.

[0193] 등록이 성숙되지 않은 경우, 서버는 등록 신호에 대해 사용자에게 프롬프트한다(예를 들어, 단계(302)). 서버는 등록이 성숙될 때까지 추가 등록 신호들(등록 발화들이 있음)에 대해 사용자에게 프롬프트한다. 일 예로서, 서버가 제1 유형의 콘텐트(예를 들어, 사용자 이름)가 포함된 등록 발화를 수신하는 경우, 서버는 사용자에게 제2 발화(예를 들어, 사용자 생일)를 프롬프트한다. 다른 예로서, 서버가 생체 인식 정보(예를 들어, 지문)를 수신하는 경우, 서버는 등록 발화를 포함하는 오디오 신호를 사용자에게 프롬프트한다.

[0194] 등록이 성숙되는 경우, 서버는 단계(310)로 진행한다. 단계(310)에서, 서버는 사용자를 위한 성문(때때로 등록자 성문이라고도 함)을 생성한다. 서버는 등록된 화자-사용자에 대한 성문을 추출하기 위해 등록 임베딩들을 통계적으로 또는 알고리즘적으로 결합한다. 일부 구현들에서, 등록 신호로부터 추출된 등록 임베딩들의 클러스터

들은 성문이 화자 클러스터의 중심을 나타내는 특정 화자(예를 들어, 화자 클러스터)와 유사한 발화 모음을 나타낸다.

[0195] 단계(312)에서, 서버는 화자 프로파일을 업데이트한다. 화자 프로파일들은 등록 단계 동안 등록하는 각 사용자(또는 화자)와 관련된 데이터 레코드들이다. 화자 프로파일에는 성문, 화자 클러스터(예를 들어, 성문과 연관된 임베딩들), 및 메타데이터가 포함된다. 도 1에 설명된 바와 같이, 메타데이터에는 발화, 등록 신호의 품질, 순음성의 전체 지속시간, 각 클러스터의 총 발화 수, 화자 특성들 및 화자-독립 특성들에 대한 정보를 포함할 수 있다. 화자 식별자들(313)은 특정 화자들을 구별하고 식별하기 위해 각 화자 프로파일과 연관된다. 도 1에 설명된 바와 같이, 화자 식별자는 임의의 개인 식별 정보를 포함하거나 나타낼 수 없다. 서버는 또한 새로운 화자 식별자를 생성하거나 콘텐트 서버에 새로운 화자 프로파일에 대한 새로운 화자 식별자를 요청할 수 있다.

[0196] 단계(314)에서, 서버는 화자 식별자를 제3자 시스템(예를 들어, 콘텐트 서버, 콜 센터)에 전송할 수 있다. 제3자 시스템은 수신된 화자 식별자를 사람 화자와 매핑할 수 있다. 예를 들어, 제3자 시스템은 룩업 테이블을 사용하여 화자 식별자를 특정 사람 화자에 매핑할 수 있다. 추가로 또는 대안으로, 제3자 시스템은 가입자 식별자 또는 기타 가구 및/또는 그룹 식별자를 각 화자 식별자와 연관시킬 수 있다. 서버는 또한 화자 식별자와 연관된 화자 프로파일(화자-독립 특성들, 화자 특성들, 메타데이터를 포함)을 전송할 수 있다.

[0197] 도 3b를 참조하면, 단계(322)에서, 서버는 생체 인식 정보를 포함하는 인바운드 신호를 수신할 수 있다. 인바운드 신호는 사용자의 이미지, 오디오 신호의 발화, 지문 등일 수 있다. 일부 구성들에서, 서버는 최종 디바이스가 사용자 입력 데이터를 능동적으로 캡처하는 것에 응답하여 인바운드 신호를 수신한다. 예를 들어, 사용자는 최종 사용자 디바이스와 능동적으로 상호작용(예를 들어, "웨이크" 단어 말하기, 버튼 누르기, 제스처 만들기)할 수 있다. 추가로 또는 대안으로, 최종 사용자 디바이스(114)는 사용자 입력 데이터를 수동적으로 캡처하며, 여기서 최종 사용자는 최종 사용자 디바이스(114)와 수동적으로 상호작용한다(예를 들어, 다른 사용자에게 말을 하는 경우, 최종 사용자 디바이스(114)는 사용자의 적극적인 조치 없이 자동으로 발화를 캡처함).

[0198] 단계(324)에서, 서버는 인바운드 신호로부터 임베딩들을 추출한다. 일부 구성들에서, 서버는 인바운드 신호에 대한 전처리를 수행(예를 들어, 인바운드 신호 분할, 인바운드 신호 스케일링, 인바운드 신호 노이즈 제거)할 수 있다. 추가로 또는 대안으로, 서버는 인바운드 신호에서 이벤트들을 식별할 수 있다. 예를 들어, 서버는 VAD 소프트웨어를 실행하여 오디오 이벤트들을 검출할 수 있다. VAD 소프트웨어는 음성과 침묵을 구별할 수 있다. 추가로 또는 대안으로, 서버는 인바운드 신호에서 객체 검출(또는 인식, 식별)을 수행할 수 있다. 예를 들어, 서버는 손과 얼굴을 인식하고 구별할 것이다. 서버는 인바운드 신호의 식별된 생체 인식 부분들(예를 들어, 음성, 지문들)을 사용하여 임베딩들을 추출할 것이다.

[0199] 인바운드 신호는 임의 명수의 화자들에 대한 임의 개수의 화자 발화들을 포함하는 오디오 레코딩을 포함하는 오디오 신호일 수 있다. 오디오 데이터에는 또한 오디오 신호와 함께 수신된 데이터 또는 메타데이터가 포함될 수 있다. 예를 들어, 오디오 데이터에는 화자 관련 정보(예를 들어, 사용자/화자 식별자, 가입자/가구 식별자, 사용자 생체 인식) 또는 통신 프로토콜들 또는 매체와 관련된 메타데이터(예를 들어, TCP/IP 헤더 데이터, 전화번호/ANI)가 포함될 수 있다.

[0200] 서버는 오디오 신호의 특징들을 추출하도록 구성된 기계 학습 모델들을 싱잉(sing)하는 임베딩들을 추출할 수 있다. 특징들에는 다양한 화자 발화들의 낮은 레벨의 스펙트럼-시간적 특징들이 포함된다. 특징들에는 또한 사용자들 또는 화자들과 관련된 다양한 데이터(예를 들어, 가입자 식별자들, 화자 식별자들, 생체 인식) 또는 프로토콜 정보 또는 데이터 패킷들로부터 추출한 메타데이터 값들(예를 들어, IP 어드레스)이 포함될 수 있다. 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기 모델)을 포함하는 기계 학습 아키텍처는 입력 오디오 신호로부터 추출된 특징들에 기초하여 추정 화자들에 대한 임베딩들을 추출한다.

[0201] 단계(326)에서, 기계 학습 아키텍처는 특정 화자에 대해 추출된 특징들 또는 임베딩들을 다른 추정 화자들의 해당 특징들 또는 임베딩들 및/또는 데이터베이스에 저장된 화자 클러스터들의 해당 특징들 또는 임베딩들과 비교하여 하나 이상의 유사도 점수들을 생성한다. 클러스터들은 화자 성문이 화자 클러스터의 중심을 나타내는 특정 화자와 유사한 발화 모음을 나타낸다. 서버는 오디오 신호에 기계 학습 아키텍처를 적용하여 특정 화자에 대한 임베딩을 추출한다. 서버는 임베딩들 또는 특징들을 화자 프로필 데이터베이스에 저장된 화자 클러스터들과 비교한 다음 화자들의 유사도 점수들을 결정한다. 각 특정 화자에 대해, 서버는 화자 임베딩(입력 오디오 신호로부터 추출됨)과 화자 프로파일 데이터베이스에 저장된 성문들 사이의 상대적 거리에 기초하여 유사도 점수 세트를 생성한다.

[0202] 추가로 또는 대안으로, 서버는 입력 오디오 신호로부터 추출된 특정 특징들에 따라 클러스터링 동작들을 수행하고 특징들에 기초하여 각 사용자에 대한 하나 이상의 클러스터링 유사도 점수들을 결정한다.

[0203] 서버는 최대 유사도 점수(327)를 갖는 화자 클러스터들과 화자들의 쌍들을 식별한다. 각 화자에 대해, 서버는 화자들 사이의 일치 가능성이 가장 높은, 특정 화자에 대해 계산된 최대 유사도 점수(327)를 출력한다.

[0204] 결정 단계(328)에서, 서버는 각 성문에 대한 최대 유사도 점수들이 하나 이상의 임계값들을 충족하는지 여부를 결정한다. 서버는 알려진 성문들 또는 예상된 특징들에 대해 해당 임베딩 또는 특징들의 유사도 점수를 평가하여 인바운드 오디오 신호에 새 화자 또는 알려진 화자에 대한 발화가 포함되는지 여부를 결정한다.

[0205] 최대 유사도 점수가 하나 이상의 임계값들을 충족하는 경우, 서버는 화자가 알려진 등록된 사용자(329)와 연관될 가능성이 있다고 결정한다. 예를 들어, 알려진 등록된 사용자(329)는 도 3a의 등록 단계 동안 등록된 사용자일 수 있다. 반면에, 서버가 최대 유사도 점수가 하나 이상의 임계값들에 미치지 못하다고 결정하는 경우, 서버는 화자가 새로운 알 수 없는 사용자(331)일 가능성이 있다고 결정한다. 서버는 특정 성문을 사용하여 새로운 화자 프로파일 및 화자 식별자를 생성할 수 있다. 화자 프로파일에는 성문, 화자 클러스터(예를 들어, 성문과 연관된 임베딩들), 화자 식별자 및 메타데이터가 포함된다.

[0206] 단계(330)에서, 서버는 (알려진 등록된 사용자(329) 및/또는 새로운 알 수 없는 사용자(331)의) 화자 식별자 및 화자 프로파일을 하나 이상의 다운스트림 애플리케이션들로 출력한다. 일부 구성들에서, 서버는 알려진 등록된 사용자(329)의 화자 프로파일에 따라 사용자를 인증한다. 추가로 또는 대안으로, 다운스트림 동작들은 화자 식별자를 사용하여 특정 화자를 식별, 인증 및/또는 인가한다. 다운스트림 애플리케이션은 화자가 알려진 등록된 사용자인지 또는 새로운 알 수 없는 사용자인지에 따라 다른 기능들을 수행할 수 있다. 예를 들어, 다운스트림 애플리케이션들은 식별된 및/또는 인증된 등록된 사용자들에 기초하여 동작들을 실행하거나, 기능을 잠금 해제하거나, 또는 동작들을 수행할 수 있다. 사용자가 새로운 알 수 없는 사용자(331)인 경우, 다운스트림 애플리케이션들은 애플리케이션의 소프트웨어, 기능 또는 정보에 대한 사용자 액세스를 제한할 수 있다.

[0207] **오디오 처리를 위한 적응적 임계화**

[0208] 도 4는 오디오 처리 시스템에서 적응적 임계화를 위한 방법(400)의 동작 단계들을 도시한다. 아래에 설명된 방법(400)은 기계 판독 가능 소프트웨어 코드를 실행하는 분석 시스템의 서버(예를 들어, 분석 서버)에 의해 수행되지만, 임의 개수의 컴퓨팅 디바이스들의 임의 개수의 프로세서들이 방법(400)의 다양한 동작들을 수행할 수 있다. 실시예들은 방법(400)에서 설명된 방법들보다 추가 동작, 적은 동작 또는 다른 동작을 포함할 수 있다. 서버는 다양한 동작들을 수행하기 위해 입력 오디오 신호들에 임의 개수의 기계 학습 모델들 및 알고리즘들을 포함하는, 기계 학습 아키텍처를 적용한다. 서버 및 기계 학습 아키텍처는 등록 단계 동안 방법(400)을 실행하지만, 서버 및 기계 학습 아키텍처는 등록 단계, 배포 단계 동안 또는 이러한 단계들의 지속적인 조합으로서 방법(400)의 다양한 동작들을 수행할 수 있다.

[0209] 단계(402)에서, 서버는 입력 오디오 신호를 수신한다. 서버는 (도 2에서와 같이) 최종 사용자 디바이스로부터 직접 입력 오디오 신호를 수신하거나 (도 1에서와 같이) 제3자 시스템(예를 들어, 콘텐트 시스템, 콜 센터 시스템)의 컴퓨팅 디바이스를 통해 입력 오디오 신호를 수신한다. 입력 오디오 신호는 화자-사용자의 발화를 포함하며, 여기서 입력 오디오 신호는 데이터 파일(예를 들어, WAV 파일, MP3 파일) 또는 데이터 스트림이다. 서버는 다른 잠재적 동작들 중에서, 오디오 신호를 음성의 세그먼트들 또는 프레임들로 파싱하거나 하나 이상의 변환 동작들(예를 들어, 고속 푸리에 변환)을 수행하는 것과 같은, 입력 오디오 신호에 대한 다양한 전처리 동작들을 수행한다.

[0210] 입력 오디오 신호는 등록 오디오 신호 또는 인바운드 오디오 신호일 수 있으며, 여기서 서버는 등록 단계 또는 배포 단계 동안 입력 오디오를 수신한다. 오디오 신호에는 또한 오디오 신호와 함께 수신된 데이터 또는 메타데이터가 포함될 수 있다. 예를 들어, 오디오 신호에는 화자 관련 정보(예를 들어, 사용자/화자 식별자, 가입자/가구 식별자, 사용자 생체 인식) 또는 통신 프로토콜들 또는 매체와 관련된 메타데이터(예를 들어, TCP/IP 헤더 데이터, 전화번호/ANI)가 포함될 수 있다.

[0211] 서버는 오디오 신호를 수신하고 오디오 신호로부터 다양한 유형의 특징들을 추출한다. 특징들에는 다양한 화자 발화들의 낮은 레벨의 스펙트럼-시간적 특징들이 포함된다. 특징들에는 또한 사용자들 또는 화자들과 관련된 다양한 데이터(예를 들어, 가입자 식별자들, 화자 식별자들, 생체 인식) 또는 프로토콜 정보 또는 데이터 패킷들로부터 추출한 메타데이터 값들(예를 들어, IP 어드레스)이 포함될 수 있다. 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기 모델)을 포함하는 기계 학습 아키텍처는 입력 오디오 신호로부터 추출된 특징들에 기초하여

추정 화자들에 대한 임베딩들을 추출한다.

[0212] 일부 실시예들에서, 서버 또는 기계 학습 아키텍처는 오디오 신호에 음성 활동 검출(VAD) 모델을 적용한다. VAD는 오디오 신호 내에서 발화들의 인스턴스들을 검출하고 검출된 발화들을 포함하는 오디오 신호의 세그먼트들을 추출하거나 아니면 식별하도록 훈련된 기계 학습 모델이다. 일부 경우에, VAD는 요약된 음성 전용 오디오 신호를 생성한다. 서버는 음성 세그먼트들 및/또는 요약된 오디오 신호를 데이터베이스(예를 들어, 화자 프로파일 데이터베이스, 성문 데이터베이스)에 또는 일부 다른 비일시적 기계 판독 가능 저장 매체에 저장할 수 있다.

[0213] 단계(404)에서, 각 특정 사용자에 대해, 서버는 임베딩 또는 성문을 분석 데이터베이스의 데이터베이스 레코드(화자 프로파일을 나타냄)에 저장한다. 서버는 특정 화자와 관련된 다양한 유형의 사용자 정보를 더 저장한다. 사용자 정보는 예를 들어 사용자에 대해 생성된 사용자별 유사도 임계값을 포함한다. 화자 프로파일이 새롭거나 아니면 사용자별 유사도 임계값이 없는 경우, 서버는 미리 구성된 기본 유사도 임계값을 화자 프로파일에 저장한다._

[0214] 결정 단계(406)에서, 서버는 성문이 성숙한지 안정적인지를 나타내는, 하나 이상의 성숙도 임계값들을 충족하는지 여부를 결정한다. 서버는 화자 프로파일 및 발화들과 관련된 하나 이상의 성숙도 요소들을 식별한다. 성숙도 인자의 비제한적인 예들에는 등록 발화들의 수, 발화 전반에 걸친 순 음성의 전체 지속시간, 및 오디오 신호들로부터의 오디오 품질이 포함된다.

[0215] 서버는 성문이 성숙한지 여부를 결정하기 위해 임의 개수의 알고리즘들을 사용할 수 있다. 예를 들어, 서버는 성숙도 인자(예를 들어, 발화들의 수)를 성숙도 인자(예를 들어, 발화들의 임계 수)에 해당하는 미리 구성된 성숙도 임계값과 비교한다. 다른 예로서, 서버는 성숙도 인자들을 통계적으로 또는 알고리즘적으로 결합하고 결합된 성숙도 인자를 결합된 성숙도 인자에 대응하는 미리 구성된 성숙도 임계값과 비교한다. 미리 구성된 오수락률은 미리 구성된 성숙도 임계값을 나타낸다. 관리자가 수동으로 입력하거나 다양한 기계 학습 모델들이 알고리즘적으로 오수락률을 결정한다. 오수락이 증가함에 따라, 성숙도 임계값이 증가하며, 이로써 서버에서 성숙도 인자가 성숙도 임계값에 미치지 못하고 성문이 충분히 성숙하지 않은 것으로 결정할 가능성이 높아진다.

[0216] 일부 실시예들에서, 서버는 계층화된 오수락률에 해당하는 계층화된 성숙도 임계값들을 사용한다. 예를 들어, 서버는 다양한 오수락률로 성숙도 임계값들(예를 들어, 임계값 스케줄)의 테이블을 저장할 수 있다. 도 12를 참조하면, 예(1200)는 단일 성숙도 인자(등록 임베딩 수) 및 다양한 오수락률에 기초한 임계값 스케줄러를 예시한다. 일부 구현들에서, 임계값 스케줄들은 유사도 점수들을 여러 성숙도 인자들(예를 들어, 등록 수(1202))과 연관시킬 수 있다. 열(1202)은 단일 성숙도 인자(예를 들어, 등록 발화의 수(1202))에 기초한 성숙도 임계값을 나타낸다. 서로 다른 미리 결정된 오수락률(FAR)이 각각 열들(1204 및 1206)에 표시된다. 열(1204)은 0.5%에서 미리 결정된 FAR에 기초한 유사도 임계값을 나타낸다. 열(1206)은 5%에서 미리 결정된 FAR에 기초한 유사도 임계값을 나타낸다. 열(1204)에 표시된 유사도 임계값들은 높은 유사도 임계값들로 간주된다. 열(1206)의 유사도 임계값들은 낮은 유사도 임계값들로 간주된다.

[0217] 도 4를 다시 참조하면, 단계(410)에서, 서버는 성문이 성숙도 임계값에 못할 정도로 성숙되지 않았다고 서버가 결정(단계(408)에서)하는 것에 응답하여 화자에 대한 유사도 임계값을 조정한다. 서버는 오수락률에 따라 유사도 임계값을 조정하고, 이에 따라 서버는 오수락률로 표시되는 원하는 정확도 레벨을 충족하도록 유사도 임계값을 높이거나 낮춘다. 예를 들어, 서버는 성숙도 인자들(예를 들어, 발화 수)이 주어진 성숙도 인자 임계값을 충족하지 못하는 경우 특정 화자에 대한 유사도 임계값을 높인다. 서버는 서버가 추후 인바운드 오디오 신호들에 대한 오수락률에 따라 성문을 평가하도록 유사도 점수를 업데이트한다. 서버는 성숙도 임계값이 충족될 때까지 추가 발화 또는 등록 임베딩을 화자에게 반복적으로 프롬프트한다.

[0218] 도 12를 다시 참조하면, 일 예에서, 성숙도 임계값은 10개의 등록 발화들로 설정될 수 있다. 서버가 성문이 성숙도 임계값을 충족하지 않는다고 결정하는 경우(예를 들어, 단계(402)에서 수신된 입력 오디오 신호가 7번째 등록 발화에 대한 것인 경우), (특정 가입자 식별자들 또는 제3자 시스템들에 대한) FAR에 따라, 화자와 관련된 높은 유사도 점수는 4.29(예를 들어, 점수 임계값(1208))이고 낮은 유사도 점수는 0.1(예를 들어, 점수 임계값(1210))이다. 서버는 화자와 관련된 유사도 점수들을 화자 데이터베이스에 저장한다. 화자가 다시 말하는 경우, 추후 인바운드 오디오 신호들의 유사도를 평가하는 데 사용되는 임계값들은 화자와 관련된 높은 유사도 점수가 4.29(예를 들어, 점수 임계값(1208))이고 낮은 유사도 점수가 0.1(예를 들어, 점수 임계값(1210))이다.

[0219] 도 4를 다시 참조하면, 단계(412)에서, 서버는 성문 및 유사도 임계값을 화자 프로파일에 저장하고 서버가 성문이 성숙하다고 결정(단계(408)에서)할 때 추후 인바운드 오디오 신호들에 대해 성문 및 유사도 임계값을 적용한

다. 서버는 특정 화자로부터의 발화를 포함한다고 알려진 추후 인바운드 오디오 신호들에 대해 성문 및 유사도 임계값을 적용한다. 일부 구현들에서, 서버는 성문 및 유사도 임계값을 화자 프로파일 데이터베이스에 저장한다. 일부 구현들에서, 서버는 성문 및 유사도 임계값을 서버와 통신하는 최종 사용자 디바이스 또는 다른 디바이스에 저장한다.

[0220] 도 12를 다시 참조하면, 상술된 예에서, 성숙도 임계값은 10개의 등록 발화들로 설정된다. 서버가 성문이 성숙도 임계값을 충족한다고 결정하는 경우(예를 들어, 단계(402)에서 수신된 입력 오디오 신호가 10번째 등록 발화에 대한 것인 경우), (특정 가입자 식별자들 또는 제3자 시스템들에 대한) FAR에 따라, 화자와 관련된 높은 유사도 점수는 4.37(예를 들어, 점수 임계값(1212))이고 낮은 유사도 점수는 0.15(예를 들어, 점수 임계값(1214))이다. 서버는 화자와 관련된 유사도 점수들을 화자 데이터베이스에 저장한다. 화자가 다시 말하는 경우, 추후 인바운드 오디오 신호들의 유사도를 평가하는 데 사용되는 임계값들은 화자와 관련된 높은 유사도 점수가 4.29(예를 들어, 점수 임계값(1208))이고 낮은 유사도 점수가 0.1(예를 들어, 점수 임계값(1210))이다.

[0221] **비지도(Unsupervised) 클러스터링**

[0222] 도 5는 오디오 처리 시 강한 발화와 약한 발화를 식별하고 평가하기 위한 방법(500)의 실행 단계들을 도시한다. 아래에 설명된 방법(500)은 기계 판독 가능 소프트웨어 코드를 실행하는 분석 시스템의 서버(예를 들어, 분석 서버)에 의해 수행되지만, 임의 개수의 컴퓨팅 디바이스들의 임의 개수의 프로세서들이 방법(500)의 다양한 동작들을 수행할 수 있다. 실시예들은 방법(500)에서 설명된 방법들보다 추가 동작, 적은 동작 또는 다른 동작을 포함할 수 있다. 서버는 다양한 동작들을 수행하기 위해 임의 개수의 기계 학습 모델들 및 알고리즘들을 포함하는, 기계 학습 아키텍처를 적용한다.

[0223] 단계(502)에서, 서버는 최종 사용자 디바이스로부터 입력 오디오 신호를 수신하고 입력 오디오 신호로부터 다양한 유형의 특징들을 추출한다. 입력 오디오 신호는 기계 판독 가능 형식의 오디오 데이터를 포함하는 데이터 파일 또는 데이터 스트림을 포함한다. 오디오 데이터는 임의 명수의 화자들에 대한 임의 개수의 화자 발화들을 포함하는 오디오 레코딩을 포함한다. 입력 오디오 신호는 화자-사용자의 발화를 포함하며, 여기서 입력 오디오 신호는 데이터 파일(예를 들어, WAV 파일, MP3 파일) 또는 데이터 스트림이다. 입력 오디오 신호는 등록 오디오 신호 또는 인바운드 오디오 신호일 수 있으며, 여기서 서버는 등록 단계 또는 배포 단계 동안 입력 오디오를 수신한다. 서버는 입력 오디오 신호로부터 스펙트럼-시간적 특징들 또는 메타데이터와 같은 다양한 유형의 특징들을 추출한다. 추가로 또는 대안으로, 서버는 다른 잠재적 동작들 중에서, 오디오 신호를 음성의 세그먼트들 또는 프레임들로 파싱하거나 하나 이상의 변환 동작들(예를 들어, 고속 푸리에 변환)을 수행하는 것과 같은, 입력 오디오 신호에 대한 다양한 전처리 동작들을 수행한다.

[0224] 단계(504)에서, 서버는 화자들에 대해 추출된 임베딩들 또는 특징들을 화자 프로파일 데이터베이스에 저장된 화자 클러스터들과 비교한 다음 화자들의 유사도 점수들을 결정한다. 클러스터들은 화자 성문이 화자 클러스터의 중심을 나타내는 특정 화자와 유사한 발화 모음을 나타낸다. 서버는 오디오 신호에 하나 이상의 기계 학습 모델들을 갖는 기계 학습 아키텍처를 적용하여 특정 화자에 대한 임베딩을 추출한다. 각 특정 화자에 대해, 서버는 화자 임베딩(입력 오디오 신호로부터 추출됨)과 화자 프로파일 데이터베이스에 저장된 성문들 사이의 상대적 거리에 기초하여 유사도 점수 세트를 생성한다.

[0225] 추가로 또는 대안으로, 서버는 입력 오디오 신호로부터 추출된 특정 특징들에 따라 클러스터링 동작들을 수행하고 특징들에 기초하여 각 사용자에 대한 하나 이상의 클러스터링 유사도 점수들을 결정한다.

[0226] 단계(506)에서, 서버는 최대 유사도 점수를 갖는 각 쌍의 화자와 클러스터를 식별한다. 각 화자에 대해, 서버는 화자들 사이의 일치 가능성이 가장 높은, 특정 화자에 대해 계산된 최대 유사도 점수를 출력한다.

[0227] 결정 단계(508)에서, 서버는 특정 화자의 임베딩(또는 특징들)에 대한 유사도 점수가 하나 이상의 유사도 임계값들을 충족하는지 여부를 결정한다. 서버는 알려진 성문들 또는 예상된 특징들에 대해 해당 임베딩 또는 특징들의 유사도 점수를 평가하여 입력 오디오 신호에 새 화자 또는 알려진 화자에 대한 발화가 포함되는지 여부를 결정한다. 서버가 특정 임베딩에 대한 유사도 점수가 유사도 임계값을 충족한다고 결정하는 경우, 서버는 마찬가지로 임베딩이 알려진 등록된 사용자와 연관될 가능성이 있다고 결정한다. 반면에, 서버가 특정 임베딩에 대한 유사도 점수가 유사도 임계값에 못 미친다고 결정하는 경우, 서버는 임베딩이 새로운 사용자일 가능성이 있다고 결정한다.

[0228] (도 6에서와 같이) 일부 실시예들에서, 서버는 특정 화자에 대해 출력된 유사도 점수를 낮은 유사도 임계값 및 높은 유사도 임계값과 비교하여 낮은 품질의 발화로 인한 품질이 낮은 임베딩들을 제어할 수 있는 세분성 레벨

을 제공한다. 이러한 실시예들에서, 서버가 특정 임베딩에 대한 유사도 점수가 높은 유사도 임계값을 충족한다고 결정하는 경우, 서버는 마찬가지로 임베딩이 알려진 등록된 사용자와 연관될 가능성이 있다고 결정한다. 반면에, 서버가 특정 임베딩에 대한 유사도 점수가 낮은 유사도 임계값에 못 미친다고 결정하는 경우, 서버는 임베딩이 새로운 사용자일 가능성이 있다고 결정한다. 하기에 논의된 바와 같이, 유사도 점수가 낮은 임계값과 높은 임계값 사이에 있는 경우, 서버는 관련 오디오 데이터 및 유사도 점수를 버퍼 메모리, 화자 프로파일 또는 기타 격리 메모리 위치에 저장한다.

[0229] 단계(510)에서, 서버는 서버가 특정 화자에 대한 임베딩이 유사도 임계값에 미치지 못했다고 결정하는 것(단계(508)에서)에 응답하여 새로운 성문과 새로운 화자 프로파일을 생성한다. 서버는 분석 데이터베이스 또는 임시 또는 게스트 화자 프로파일들을 저장하도록 구성된 다른 메모리 위치에 새 화자 프로파일을 생성한다. 서버는 다양한 유형의 데이터를 새 화자 프로파일에 저장하며, 이는 서버가 입력 오디오 신호(예를 들어, 화자 식별자, 임베딩들, 성문, 특징들, 메타데이터, 디바이스 정보)로부터 식별하거나 추출한다.

[0230] 단계(512)에서, 서버는 서버가 특정 알려진 화자에 대한 임베딩이 유사도 임계값을 충족한다고 결정하는 것(단계(508)에서)에 응답하여 데이터베이스에서 저장된 성문 및 기존 등록된(또는 알려진) 화자 프로파일을 업데이트한다. 서버는 분석 데이터베이스에서의 또는 임시 또는 게스트 화자 프로파일에서의 알려진 화자 프로파일에 대해 알려진 화자 프로파일을 업데이트한다. 서버는 다양한 유형의 데이터를 알려진 화자 프로파일에 저장하며, 이는 서버가 입력 오디오 신호(예를 들어, 화자 식별자, 임베딩들, 성문, 특징들, 메타데이터, 디바이스 정보)로부터 식별하거나 추출한다.

[0231] 옵션 단계(514)에서, 서버는 화자 클러스터들을 재평가하고 화자 프로파일 데이터베이스의 화자 프로파일들을 업데이트하기 위해 하나 이상의 재클러스터링 동작들을 수행한다. 서버는 특정 트리거링 조건에 응답하여 재클러스터링 동작들을 수행한다. 트리거링 조건의 비제한적인 예들은 미리 구성된 주기적 시간 간격 또는 서버가 그 중에서도 가입자(예를 들어, 가구) 또는 사용자와 관련된 발화들의 임계 수를 수신할 때를 포함할 수 있다. 서버는 새 성문들을 생성하거나 화자 프로파일들에 대한 기존 성문들을 업데이트하기 위해 화자들에 대한 특징들 또는 임베딩들을 추출한다. 서버는 각각의 특정 발화와 각각의 특정 성문 또는 다른 유형의 클러스터 중심에 대해 추출된 특징들 또는 임베딩들 사이의 상대적 거리들에 기초하여 화자들에 대한 유사도 점수들을 재계산한다. 서버는 새롭거나 업데이트된 성문들을 화자와 관련된 다양한 유형의 데이터와 함께 새롭거나 업데이트된 화자 프로파일들에 저장한다.

[0232] 일부 실시예들에서, 서버에 의해 수행되는 재클러스터링 동작은 계층적 클러스터링 동작이다(도 7에서와 같이 참조). 계층적 클러스터링은 작은 누적 식별/인증 오류들이 발생할 가능성을 최소화하여, 화자 클러스터들의 순도를 높인다. 서버는 성문들 및 식별자들 사이의 상대적 거리들 또는 기타 비교 차이들 또는 클러스터링 알고리즘들(예를 들어, PLDA, 코사인 거리)을 결정하고 유사도 점수 임계값을 충족하는 기존 또는 새로운 성문들을 식별한다.

[0233] **오디오 처리를 위한 강하고 약한 발화 식별**

[0234] 도 6은 오디오 처리 시 화자들을 클러스터링하기 위한 방법(600)의 동작 단계들을 도시한다. 아래에 설명된 방법(600)은 기계 판독 가능 소프트웨어 코드를 실행하는 분석 시스템의 서버(예를 들어, 분석 서버)에 의해 수행되지만, 임의 개수의 컴퓨팅 디바이스들의 임의 개수의 프로세서들이 방법(600)의 다양한 동작들을 수행할 수 있다. 실시예들은 방법(600)에서 설명된 방법들보다 추가 동작, 적은 동작 또는 다른 동작을 포함할 수 있다. 서버는 다양한 동작들을 수행하기 위해 입력 오디오 신호들에 임의 개수의 기계 학습 모델들 및 알고리즘들을 포함하는, 기계 학습 아키텍처를 적용한다. 서버 및 기계 학습 아키텍처는 등록 단계 동안 방법(600)을 실행하지만, 서버 및 기계 학습 아키텍처는 등록 단계, 배포 단계 동안 또는 이러한 단계들의 지속적인 조합으로서 방법(600)의 다양한 동작들을 수행할 수 있다.

[0235] 서버는 (능동 또는 수동) 등록 동작들, 배포 동작들 동안 및/또는 데이터베이스 업데이팅 동작들을 재클러스터링하는 동안 방법(600)을 실행한다. 클러스터링 동작들의 경우, 서버는 현재 오디오 신호들 또는 이력 오디오 신호들에 훈련된 기계 학습 아키텍처의 동작들을 적용하며, 여기서 오디오 신호들에는 등록 오디오 신호들, 인바운드 오디오 신호들(배포 단계 동안 수신됨) 또는 저장된 오디오 신호들이 포함될 수 있다. 기계 학습 아키텍처에는 전처리(예를 들어, 특징 추출) 및 클러스터링 동작들을 포함하여, 서버가 특정 오디오 신호들에 적용하는 임의 개수의 기계 학습 모델들과 다양한 기타 동작들이 포함된다. 클러스터링 동작들은 오디오 신호들의 화자 발화들을 사용하여 새롭거나 알려진 화자 인식을 용이하게 한다. 서버는 오디오 신호로부터 특징들 또는 특징 벡터들(예를 들어, 임베딩들)을 추출한 다음 추출된 정보(예를 들어, 특징들, 임베딩들)를 오디오 신호에 있

는 화자들에 해당하는 클러스터들로 클러스터링한다. 방법(600)은 비지도 클러스터링 동작들을 포함하지만, 일부 실시예들에서, 서버는 지도(supervised) 클러스터링 동작들을 수행할 수 있다.

[0236] 단계(602)에서, 서버는 최종 사용자 디바이스로부터 입력 오디오 신호를 수신하고 입력 오디오 신호로부터 다양한 유형의 특징들을 추출한다. 오디오 신호는 기계 판독 가능 형식의 오디오 데이터를 포함하는 데이터 파일 또는 데이터 스트림일 수 있다. 오디오 데이터는 임의 명수의 화자들에 대한 임의 개수의 화자 발화들을 포함하는 오디오 레코딩을 포함한다. 오디오 데이터에는 또한 오디오 신호와 함께 수신된 데이터 또는 메타데이터가 포함될 수 있다. 예를 들어, 오디오 데이터에는 화자 관련 정보(예를 들어, 사용자/화자 식별자, 가입자/가구 식별자, 사용자 생체 인식) 또는 통신 프로토콜들 또는 매체와 관련된 메타데이터(예를 들어, TCP/IP 헤더 데이터, 전화번호/ANI)가 포함될 수 있다.

[0237] 서버는 오디오 신호를 수신하고 오디오 데이터로부터 다양한 유형의 특징들을 추출한다. 특징들에는 다양한 화자 발화들의 낮은 레벨의 스펙트럼-시간적 특징들이 포함된다. 특징들에는 또한 사용자들 또는 화자들과 관련된 다양한 데이터(예를 들어, 가입자 식별자들, 화자 식별자들, 생체 인식) 또는 프로토콜 정보 또는 데이터 패킷들로부터 추출한 메타데이터 값들(예를 들어, IP 어드레스)이 포함될 수 있다. 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기 모델)을 포함하는 기계 학습 아키텍처는 입력 오디오 신호로부터 추출된 특징들에 기초하여 추정 화자들에 대한 임베딩들을 추출한다.

[0238] 단계(604)에서, 기계 학습 아키텍처는 특정 화자에 대해 추출된 특징들 또는 임베딩들을 다른 추정 화자들의 해당 특징들 또는 임베딩들 및/또는 데이터베이스에 저장된 화자 클러스터들의 해당 특징들 또는 임베딩들과 비교하여 하나 이상의 유사도 점수들을 생성한다. 클러스터들은 화자 성문이 화자 클러스터의 중심을 나타내는 특정 화자와 유사한 발화 모음을 나타낸다. 서버는 오디오 신호에 기계 학습 아키텍처를 적용하여 특정 화자에 대한 임베딩을 추출한다. 서버는 임베딩들 또는 특징들을 화자 프로필 데이터베이스에 저장된 화자 클러스터들과 비교한 다음 화자들의 유사도 점수들을 결정한다. 각 특정 화자에 대해, 서버는 화자 임베딩(입력 오디오 신호로부터 추출됨)과 화자 프로파일 데이터베이스에 저장된 성문들 사이의 상대적 거리에 기초하여 유사도 점수 세트를 생성한다.

[0239] 추가로 또는 대안으로, 서버는 입력 오디오 신호로부터 추출된 특정 특징들에 따라 클러스터링 동작들을 수행하고 추출된 특징들에 기초하여 각 사용자에 대한 하나 이상의 클러스터링 유사도 점수들을 결정한다.

[0240] 서버는 최대 유사도 점수를 갖는 화자 클러스터들과 화자들의 쌍들을 식별한다. 각 화자에 대해, 서버는 화자들 사이의 일치 가능성이 가장 높은, 특정 화자에 대해 계산된 최대 유사도 점수를 출력한다.

[0241] 단계(606)에서, 서버는 특정 화자에 대한 유사도 점수가 낮은 유사도 임계값 또는 높은 유사도 임계값을 충족하는지 여부를 결정한다. 서버는 알려진 성문들 또는 예상된 특징들에 대해 해당 임베딩 또는 특징들의 유사도 점수를 평가하여 입력 오디오 신호에 새 화자 또는 알려진 화자에 대한 발화가 포함되는지 여부를 결정한다. 서버가 특정 화자 임베딩에 대한 유사도 점수가 높은 유사도 임계값을 충족한다고 결정하는 경우, 서버는 화자가 알려진 등록된 사용자와 연관될 가능성이 있다고 결정한다. 반면에, 서버가 특정 임베딩에 대한 유사도 점수가 낮은 유사도 임계값에 못 미친다고 결정하는 경우, 서버는 화자가 새로운 알 수 없는 사용자일 가능성이 있다고 결정한다. 서버는 유사도 점수가 낮은 임계값을 충족하지만 높은 임계값도 충족하지 못한다고 결정하는 경우, 서버는 다음을 결정한다.

[0242] 단계(608)에서, 서버는 서버가 특정 화자에 대한 임베딩이 낮은 유사도 임계값에 미치지 못했다고 결정하는 것(단계(606)에서)에 응답하여 새로운 성문과 새로운 화자 프로파일을 생성한다. 서버는 분석 데이터베이스 또는 임시 또는 게스트 화자 프로파일들을 저장하도록 구성된 다른 메모리 위치에 새 화자 프로파일을 생성한다. 서버는 다양한 유형의 데이터를 새 화자 프로파일에 저장하며, 이는 서버가 입력 오디오 신호(예를 들어, 화자 식별자, 임베딩들, 성문, 특징들, 메타데이터, 디바이스 정보)로부터 식별하거나 추출한다.

[0243] 단계(610)에서, 서버는 서버가 임베딩이 낮은 임계값을 충족하지만 높은 임계값을 충족하지 못한다고 결정하는 것(단계(606)에서)에 응답하여 알려진 사용자에 대한 약한 임베딩 목록을 생성하거나 업데이트한다. 약한 임베딩 목록은 특정 화자와 관련된 버퍼 또는 격리 저장 위치로 동작하지만, 임베딩(및 발화)은 약한 임베딩을 특정 화자 클러스터의 일부로 바로 특성화하기 위한 해당 성문과의 충분한 유사도가 부족하다. 약한 임베딩은 오디오 데이터 및 알려진 사용자로부터 잠재적으로 발생한 다양한 유형의 데이터(예를 들어, 오디오 레코딩들, 발화들, 메타데이터, 임베딩)와 함께 저장된다.

[0244] 단계(612)에서, 서버는 서버가 특정 알려진 화자에 대한 임베딩이 유사도 임계값을 충족한다고 결정하는 것(단

계(606)에서)에 응답하여 데이터베이스에서 저장된 성문 및 기존 등록된(또는 알려진) 화자 프로파일을 업데이트한다. 서버는 알려진 사용자와 관련된 강한 임베딩 목록을 업데이트한다. 강한 임베딩 목록에는 서버가 알려진 사용자에 대한 성문 또는 클러스터를 생성하는 데 사용하는 임베딩들이 포함된다. 서버는 분석 데이터베이스에서의 또는 임시 또는 게스트 화자 프로파일에서의 알려진 화자에 대해 화자 프로파일을 업데이트한다. 서버는 강한 임베딩으로 수신된 다양한 유형의 데이터를 알려진 화자 프로파일에 저장하며, 이는 서버가 입력 오디오 신호(예를 들어, 화자 식별자, 임베딩들, 성문, 특징들, 메타데이터, 디바이스 정보)로부터 식별하거나 추출한다.

[0245] 옵션 단계(614)에서, 서버는 재클러스터링 동작을 수행하고, 데이터베이스의 클러스터들을 업데이트하며, 그에 따라 데이터베이스를 업데이트한다. 서버는 화자 클러스터들을 재평가하고 화자 프로파일 데이터베이스의 화자 프로파일들을 업데이트하기 위해 하나 이상의 재클러스터링 동작들을 수행한다. 서버는 특정 트리거링 조건에 응답하여 재클러스터링 동작들을 수행한다. 트리거링 조건의 비제한적인 예들은 미리 구성된 주기적 시간 간격 또는 서버가 그 중에서도 가입자(예를 들어, 가구) 또는 사용자와 관련된 발화들의 임계 수를 수신할 때를 포함할 수 있다. 서버는 새 성문들을 생성하거나 화자 프로파일들에 대한 기존 성문들을 업데이트하기 위해 화자들에 대한 특징들 또는 임베딩들을 추출한다. 서버는 각각의 특정 발화와 각각의 특정 성문 또는 다른 유형의 클러스터 중심에 대해 추출된 특징들 또는 임베딩들 사이의 상대적 거리들에 기초하여 화자들에 대한 유사도 점수들을 재계산한다. 서버는 새롭거나 업데이트된 성문들을 화자와 관련된 다양한 유형의 데이터와 함께 새롭거나 업데이트된 화자 프로파일들에 저장한다.

[0246] 일부 경우에, 서버는 약한 임베딩들이 특정 알려진 화자 또는 임의의 다른 화자와 충분히 유사한지 여부를 결정하기 위해 약한 임베딩들의 각 목록을 재평가한다. 재클러스터링은 하나 이상의 성문들, 클러스터들 또는 임계값들을 업데이트할 수 있다. 결과적으로, 서버의 재계산된 유사도 점수들에 따라, 하나 이상의 약한 임베딩들이 특정 성문들 또는 클러스터들과 더 잘 일치할 수 있다. 서버가 특정 약한 임베딩이 특정 성문의 유사도 임계값을 충족한다고 결정하는 경우, 서버는 약한 임베딩 및 관련 오디오 데이터를 특정 성문에 해당하는 화자 프로파일에 추가하고 약한 임베딩에 따라 성문 및 화자 프로파일을 업데이트한다.

[0247] 일부 경우에, 서버는 강한 임베딩들이 특정 알려진 화자 또는 임의의 다른 화자와 충분히 유사한 상태로 남아 있는지 여부를 결정하기 위해 강한 임베딩들의 각 목록을 재평가한다. 재클러스터링 동작들의 결과, 하나 이상의 강한 임베딩들이 더 이상 특정 화자 성문 또는 클러스터와 충분히 일치하지 않거나, 서버의 재계산된 유사도 점수들에 따라 다른 성문 또는 클러스터와 더 잘 일치할 수 있다. 서버가 특정 강한 임베딩이 특정 성문의 유사도 임계값을 더 이상 충족하지 않는다고 결정하는 경우, 서버는 특정 성문에 해당하는 화자 프로파일로부터 강한 임베딩 및 관련 오디오 데이터를 제거하고 남아 있는 강한 임베딩들에 따라 성문 및 화자 프로파일을 업데이트한다. 서버가 특정 강한 임베딩이 특정 성문의 유사도 임계값을 충족한다고 결정하는 경우, 서버는 강한 임베딩 및 관련 오디오 데이터를 특정 성문에 해당하는 화자 프로파일에 추가하고 강한 임베딩에 따라 성문 및 화자 프로파일을 업데이트한다.

[0248] **라벨 정정**

[0249] 도 7a는 현재 및/또는 이력 정보에 따라 하나 이상의 성문들의 라벨 식별자들(예를 들어, 화자 식별자들, 가입자 식별자들)을 정정하기 위한 방법(700a)의 동작 단계들을 도시한다. 아래에 설명된 방법(700a)은 기계 판독 가능 소프트웨어 코드를 실행하는 분석 시스템의 서버(예를 들어, 분석 서버)에 의해 수행되지만, 임의 개수의 컴퓨팅 디바이스들의 임의 개수의 프로세서들이 방법(700a)의 다양한 동작들을 수행할 수 있다. 실시예들은 방법(700a)에서 설명된 방법들보다 추가 동작, 적은 동작 또는 다른 동작을 포함할 수 있다. 서버는 다양한 동작들을 수행하기 위해 입력 오디오 신호들에 임의 개수의 기계 학습 모델들 및 알고리즘들을 포함하는, 기계 학습 아키텍처를 적용한다. 서버 및 기계 학습 아키텍처는 등록 단계 동안 방법(700a)을 실행하지만, 서버 및 기계 학습 아키텍처는 등록 단계, 배포 단계 동안 또는 이러한 단계들의 지속적인 조합으로서 방법(700a)의 다양한 동작들을 수행할 수 있다.

[0250] 서버는 최종 사용자 디바이스로부터 오디오 신호를 수신하고 오디오 신호로부터 다양한 유형의 특징들을 추출한다. 오디오 데이터는 임의 명수의 화자들에 대한 임의 개수의 화자 발화들을 포함하는 오디오 레코딩을 포함한다. 서버는 오디오 신호로부터 다양한 유형의 특징들을 추출한다. 특징들에는 다양한 화자 발화들의 낮은 레벨의 스펙트럼-시간적 특징들이 포함된다. 특징들에는 또한 사용자들 또는 화자들과 관련된 다양한 데이터(예를 들어, 가입자 식별자들, 화자 식별자들, 생체 인식) 또는 프로토콜 정보 또는 데이터 패킷들로부터 추출한 메타데이터 값들(예를 들어, IP 어드레스)이 포함될 수 있다. 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기

모델)을 포함하는 기계 학습 아키텍처는 오디오 신호로부터 추출된 특징들에 기초하여 인바운드 화자들에 대한 임베딩들을 추출한다.

[0251] 서버는 특정 화자에 대해 추출된 특징들 또는 임베딩들을 다른 추정 화자들의 해당 특징들 또는 임베딩들 및/또는 데이터베이스에 저장된 성문들의 해당 특징들 또는 임베딩들과 비교하여 하나 이상의 유사도 점수들을 생성한다. 서버는 임베딩들과 성문들 사이의 상대적 거리를 평가하여 유사도 점수를 결정한다. 서버는 유클리드 거리 측정들 및/또는 상관 기반 측정들과 같은 거리 측정들에 따라 상대적 거리를 결정할 수 있다. 추가로 또는 대안으로, 서버는 예를 들어 코사인 유사도 접근 방식들 또는 확률적 선형 판별 분석(PLDA)을 사용하여 유사도 점수를 결정함으로써 임베딩들 및 성문들의 유사도를 평가할 수 있다. 각 임베딩과 관련된 최대 유사도 점수들이 하나 이상의 임계값들을 충족하지 않는 경우, 서버는 임베딩을 사용하여 새 화자 프로파일을 생성(예를 들어, 새 클러스터, 새 성문, 새 화자 식별자로 새 화자를 식별)할 수 있다.

[0252] 도 7b는 특정 화자에 대한 클러스터들 및 기타 추정 화자들의 클러스터들을 사용하는 라벨 정정의 일 예(700b)를 예시한다. 상기에 설명된 바와 같이, 일 예에서, 임베딩과 관련된 최대 유사도 점수가 하나 이상의 임계값들을 충족하지 않는 경우 서버는 새 화자 프로파일을 생성한다. 클러스터(721), 클러스터(723), 클러스터(725), 클러스터(727), 클러스터(729)(총칭하여 "클러스터들(720)"이라 함)는 오디오 신호로부터 추출된 임베딩 클러스터들을 나타낸다. 서버는 클러스터들(720) 중 어느 하나를 화자 식별자들과 연관시키지 않을 수 있다. 추가로 또는 대안으로, 서버는 새로운 화자 식별자들을 클러스터들(720)과 연관시키기로 결정할 수 있다.

[0253] 도 7a를 다시 참조하면, 단계(702)에서, 서버는 라벨 식별자들(예를 들어, 가입자 식별자, 화자 식별자) 및 성문들을 얻는다. 서버는 데이터베이스로부터 라벨 식별자를 얻거나 알 수 없는 화자들에 대한 새 라벨 식별자들을 생성한다. 서버는 새로 생성된 화자 프로파일로부터 하나 이상의 새 라벨 식별자들과 성문들을 얻는다. 추가로 또는 대안으로, 서버는 화자 데이터베이스에 쿼리하고 특정 화자 프로파일들에 대한 데이터를 검색하여 이전 라벨 식별자들과 성문들을 얻는다. 서버는 예를 들어 가입자 식별자와 관련된 화자 프로파일들, 서버에 의해 최근에 액세스 및/또는 수정된 화자 프로파일들(예를 들어, 지난 2일 동안 분석 서버에 의해 검색된 화자 프로파일들), 특정 화자 특성들과 관련된 화자 프로파일들, 특정 화자 독립 특성들과 관련된 화자 프로파일들 및/또는 다른 메타데이터와 관련된 화자 프로파일들을 검색한다. 검색된 화자 프로파일들에는 특정 라벨 식별자들(예를 들어, 가입자 식별자, 화자 식별자) 및 성문들이 포함될 수 있다. 서버는 하나 이상의 데이터베이스들로부터 모든 화자 프로파일들을 가져올 수도 있다.

[0254] 도 7b를 다시 참조하면, 일 예로서, 서버가 특정 가입자 식별자와 관련된 데이터베이스 레코드들(예를 들어, 화자 프로파일들)에 대해 재클러스터링 동작을 실행할 때, 서버는 특정 가입자 식별자와 관련된 이전 라벨 식별자들 및 성문들을 검색한다. 서버는 오디오 신호로부터 추출된 클러스터들(720)과 연관된 가입자 식별자를 위한 데이터베이스에 쿼리하는 것에 응답하여 알려진 클러스터(731), 알려진 클러스터(733), 알려진 클러스터(735) 및 알려진 클러스터(737)(총칭하여 "알려진 클러스터들(730)"이라고 함)를 수신한다.

[0255] 도 7a를 다시 참조하면, 단계(704)에서, 서버는 다양한 성문들 중에서 쌍별 유사도들을 계산하여 성문-쌍(voiceprint-pair) 유사도 점수들을 생성한다. 서버는 각 성문을 다른 성문들 각각과 비교하여 성문-쌍 유사도 점수들을 계산한다. 서버는 예를 들어, 쌍의 두 성문들 사이의 상대 거리, 두 성문들의 코사인 유사도, 또는 두 성문들의 PLDA를 평가하여 각각의 특정 성문-쌍 유사도 점수를 계산한다. 서버는 가장 일치하는 성문-쌍들을 식별하기 위해 성문들 각각에 대한 각각의 최대 성문-쌍 유사도 점수를 식별한다.

[0256] 도 7b를 다시 참조하면, 성문 유사도 점수 매트릭스(738)는 공지된 클러스터들(730) 각각과 비교하여 클러스터들(720) 각각의 유사도 점수들을 식별한다. 유사도 점수 매트릭스(738)의 각 셀은 알려진 클러스터들(730)의 알려진 클러스터와 클러스터들(720)의 클러스터를 비교하는 유사도 점수를 나타낸다. 클러스터들(720) 및 알려진 클러스터들(730) 각각에 대한 최대 클러스터-쌍 유사도 점수는 736에서 식별된다.

[0257] 도 7a를 다시 참조하면, 단계(706)에서, 서버는 각각의 특정 최대 성문-쌍 유사도 점수를 식별한다. 선택적으로, 서버는 각각의 특정 최대 성문-쌍 유사도 점수가 미리 구성된 재-라벨(re-label) 임계값(때때로 마이그레이션 임계값이라고 함)을 충족하는지 여부를 결정할 수 있다. 재-라벨 임계값은 인바운드 임베딩들을 특정 성문과 매칭하기 위해 사용되는 본원에서 설명된 유사도 임계값과 동일할 수 있다.

[0258] 일부 구현들에서, 성문들은 성문-쌍에서 주어진 성문에 대한 성숙도 레벨에 따르는 다양한 재-라벨 임계값들과 연관될 수 있다. 예를 들어, 특정 성문은 화자 프로파일의 낮은 유사도 임계값 및 높은 유사도 임계값과 연관된다. 일 예에서, 서버는 최대 성문-쌍 유사도 점수를 예를 들어 낮은 유사도 임계값과 비교한다. 서버는 또한 임

계값들(예를 들어, 성문들 각각과 연관된 임계값들)을 통계적으로 또는 알고리즘적으로 결합할 수 있다.

[0259] 단계(708)에서, 서버는 라벨 식별자들을 재-라벨링하거나 마이그레이션한다. 서버는 성문과 연관된 라벨 식별자들을 클러스터들과 연관된 신규 또는 업데이트된 화자 프로파일들로 재-라벨링하거나 마이그레이션할 수 있다. 라벨 식별자들을 재-라벨링하는 것 외에, 서버는 성문들 각각을 평균내거나 기계 학습 아키텍처를 성문들에 적용하여 성문들을 알고리즘적으로 결합함으로써 성문들을 병합하도록 결정할 수 있다.

[0260] 도 7b를 다시 참조하면, 서버는 알려진 클러스터들(730)의 라벨 식별자들을 마이그레이션하고 클러스터들(720)의 라벨 식별자들을 대체함으로써 클러스터들(720)의 라벨 식별자들을 업데이트하거나 정정한다. 예를 들어, 클러스터(721)는 알려진 클러스터(733)와 가장 유사하며, 최대 유사도 점수 54.2가 클러스터들(721) 및 알려진 클러스터들(730)의 다른 클러스터들(예를 들어, 각각 0.3, 3.1 및 -5.9)의 유사도 점수들보다 큰 것으로 표시된다.

[0261] 마찬가지로, 클러스터(723)는 알려진 클러스터(735)와 가장 유사하며, 최대 유사도 점수 39.8이 클러스터들(723) 및 알려진 클러스터들(730)의 다른 클러스터들(예를 들어, 각각 -0.5, 2.5 및 5.4)의 유사도 점수들보다 큰 것으로 표시된다. 클러스터(723) 및 알려진 클러스터(735)와 관련된 최대 유사도 점수 39.8이 클러스터(721) 및 알려진 클러스터(733)와 관련된 최대 유사도 점수 54.2 미만이지만, 서버는 여전히 알려진(735) 라벨 식별자를 클러스터(723)로 마이그레이션했다. 클러스터(721), 클러스터(723), 클러스터(725), 클러스터(727)와 연관된 라벨 식별자들의 서버의 재-라벨 또는 마이그레이션은 클러스터들(클러스터(721), 클러스터(723), 클러스터(725), 클러스터(727))이 이전에 서버에 의해 식별되었음을 나타낸다(예를 들어, 알려진(731), 알려진(733), 알려진(735) 및 알려진(737)). 클러스터(729)가 추정 화자 프로파일들 중 어느 하나와 충분히 유사하지 않았기 때문에 서버는 클러스터(729)에 대한 새 화자 프로파일을 만들었다. 서버는 새로운 화자 프로파일을 생성하는 서버를 나타내기 위해 클러스터(729)에 대해 새로운 라벨 식별자(Known(740))를 생성했다.

[0262] **완전 수동적 및 지속적 등록**

[0263] 도 8은 수동적 및 지속적 등록 구성들을 사용하여 오디오 처리를 위한 방법(800)의 동작 단계들을 도시한다. 아래에 설명된 방법(800)은 기계 판독 가능 소프트웨어 코드를 실행하는 분석 시스템의 서버(예를 들어, 분석 서버)에 의해 수행되지만, 임의 개수의 컴퓨팅 디바이스들의 임의 개수의 프로세서들이 방법(800)의 다양한 동작들을 수행할 수 있다. 실시예들은 방법(800)에서 설명된 방법들보다 추가 동작, 적은 동작 또는 다른 동작을 포함할 수 있다. 서버는 다양한 동작들을 수행하기 위해 입력 오디오 신호들에 임의 개수의 기계 학습 모델들 및 알고리즘들을 포함하는, 기계 학습 아키텍처를 적용한다. 서버 및 기계 학습 아키텍처는 등록 단계 동안 방법(800)을 실행하지만, 서버 및 기계 학습 아키텍처는 등록 단계, 배포 단계 동안 또는 이러한 단계들의 지속적인 조합으로서 방법(800)의 다양한 동작들을 수행할 수 있다.

[0264] 단계(802)에서, 서버는 하나 이상의 화자들의 하나 이상의 발화들을 포함하는 입력 오디오 신호를 수신한다. 서버는 최종 사용자 디바이스로부터 입력 오디오 신호를 수신하고 입력 오디오 신호로부터 다양한 유형의 특징들을 추출하며, 여기서 서버는 최종 사용자 디바이스로부터 직접 또는 중개 디바이스(예를 들어, 제3자 서버)를 통해 입력 오디오 신호를 수신한다. 입력 오디오 신호는 기계 판독 가능 형식의 오디오 데이터를 포함하는 데이터 파일 또는 데이터 스트림을 포함한다. 오디오 데이터는 임의 명수의 화자들에 대한 임의 개수의 화자 발화들을 포함하는 오디오 레코딩을 포함한다. 입력 오디오 신호는 화자-사용자의 발화를 포함하며, 여기서 입력 오디오 신호는 데이터 파일(예를 들어, WAV 파일, MP3 파일) 또는 데이터 스트림이다. 입력 오디오 신호는 등록 오디오 신호 또는 인바운드 오디오 신호일 수 있으며, 여기서 서버는 등록 단계 또는 배포 단계 동안 입력 오디오를 수신한다. 서버는 입력 오디오 신호로부터 스펙트럼-시간적 특징들 또는 메타데이터와 같은 다양한 유형의 특징들을 추출한다. 추가로 또는 대안으로, 서버는 다른 잠재적 동작들 중에서, 오디오 신호를 음성의 세그먼트들 또는 프레임들로 파싱하거나 하나 이상의 변환 동작들(예를 들어, 고속 푸리에 변환)을 수행하는 것과 같은, 입력 오디오 신호에 대한 다양한 전처리 동작들을 수행한다.

[0265] 단계(804)에서, 서버는 입력 오디오 신호로부터 추출된 특징들에 대해 임의 개수의 기계 학습 모델들을 포함하는 기계 학습 아키텍처를 적용한다. 기계 학습 아키텍처의 임베딩 추출 모델은 입력 오디오 신호로부터 추출된 특징들을 사용하여 인바운드 화자에 대한 인바운드 임베딩을 추출한다.

[0266] 결정 단계(806)에서, 서버는 데이터베이스(예를 들어, 성문 데이터베이스, 화자 프로파일 데이터베이스)가 비어 있는지 여부를 결정한다. 데이터베이스에는 제3자 콘텐트 서비스 또는 데이터 분석 서비스의 고객들인 특정 가구들, 가입자들 또는 기타 개인들의 집단에 대한 데이터 레코드들이 포함된다. 데이터 레코드들에는 화자 식별

자들과 연관된 특정 화자들에 대한 화자 프로파일들이 포함된다. 일부 구현들에서, 서버는 데이터베이스의 일부가 비어 있는지 여부를 결정한다. 예를 들어, 서버는 데이터베이스에 특정 가입자 식별자와 연관된 임의의 화자 프로파일들(예를 들어, 가구의 화자 프로파일들)이 포함되어 있는지 여부를 결정한다.

[0267] 단계(808)에서, 서버는 서버가 데이터베이스가 비어 있지 않다고 결정할 때(단계(806(에서), 데이터베이스에 저장된 각 성문과 각 임베딩 사이의 상대적 거리에 기초하여 발화들에 대해 생성된 화자 임베딩에 대한 유사도 점수들을 생성한다. 각 특정 화자에 대해, 서버는 특정 성문에 (유사도 점수에 의해) 가장 근접하게 일치하는 특정 인바운드 화자를 나타내는 최대 유사도 점수(809)를 출력한다.

[0268] 결정 단계(812)에서, 각 특정 인바운드 화자 임베딩에 대해, 서버는 대응하는 최대 유사도 점수(809)가 낮은 유사도 임계값을 충족하는지 여부를 결정한다. 일부 경우에, 낮은 유사도 임계값은 미리 구성된 기본값 또는 특정 성문 및 추정 등록된 화자에 맞게 조정된 적응 임계값이다(도 4에서와 같이).

[0269] 결정 단계(814)에서, 서버가 특정 최대 유사도 점수(809)가 낮은 임계값을 충족한다고 결정하는 경우(단계(812)에서), 서버는 최대 유사도 점수(809)가 높은 유사도 임계값을 충족하는지 여부를 결정한다. 일부 경우에, 높은 유사도 임계값은 미리 구성된 기본값 또는 특정 성문 및 추정 등록된 화자에 맞게 조정된 적응 임계값이다(도 4에서와 같이).

[0270] 단계(816)에서, 서버가 최대 유사도 점수(809)가 높은 임계값을 충족한다고 결정(단계(814)에서)하는 경우, 강한 임베딩 목록을 업데이트한다. 서버는 성문들을 생성하기 위해 강한 임베딩들을 사용한다. 예를 들어, 서버는 특정 인바운드 강한 임베딩을 사용하여 특정 성문을 업데이트한다. 서버는 업데이트된 성문을 포함하도록 해당 화자 프로파일을 추가로 업데이트할 수 있다. 특정 임베딩에 대한 최대 유사도 점수(809)가 높은 임계값을 충족했기 때문에, 서버는 특정 인바운드 임베딩이 추정 등록된 화자일 가능성이 있다고 결정한다.

[0271] 단계(822)에서, 강한 임베딩 목록을 업데이트한 후, 서버는 강한 인바운드 임베딩을 포함하도록 화자 프로파일들이 포함된 데이터베이스를 업데이트한다. 서버는 인바운드 임베딩과 가장 잘 일치하는 성문이 있는 특정 화자 프로파일에 인바운드 임베딩을 추가한다. 서버는 인바운드 임베딩에 가장 잘 일치하는 성문과 연관된 화자 식별자를 가진 화자 프로파일을 업데이트한다.

[0272] 단계(820)에서, 서버는 서버가 데이터베이스가 비어 있다고 결정(단계(806)에서)하거나 특정 인바운드 임베딩에 대한 최대 유사도 점수(809)가 낮은 유사도 임계값을 충족하지 못한다고 결정(단계(812)에서)할 때, 데이터베이스에 새로운 화자 프로파일을 생성한다. 서버는 새 화자 식별자를 새 화자 프로파일에 할당하며, 여기서 서버는 제3자 서버로부터 화자 식별자를 수신한다. 일부 구현들에서, 서버가 제3자 서버에 의해 사용되는 해당 화자 식별자의 해시된(또는 난독화된) 버전을 수신하고, 이로써 서버가 특정 화자에 대한 임의의 개인 식별 정보를 수신하지 못하도록 하여 화자 프라이버시(privacy)를 유지한다.

[0273] 일부 경우에, 새 화자 프로파일은 미리 결정된 제한된 수명 주기를 가진 임시 또는 게스트 화자 프로파일이다. 이러한 경우에, 서버 또는 데이터베이스는 임시 프로파일을 유지 관리하기 위해 미리 구성된 양의 시간 후에 데이터베이스로부터 임시 프로파일의 데이터를 퍼징한다. 서버 또는 데이터베이스는 또한 서버가 (단계(812) 또는 단계(814)에서와 같이) 임계값 임시 성문과 일치시키기 위한 높거나 낮은 임계값을 충족하는 다른 인바운드 임베딩을 식별하는 각 인스턴스에 대해, 임시 프로파일을 유지하기 위해 이 수명 주기 클록을 다시 시작한다.

[0274] 서버가 임시 프로파일의 성문이 성숙한 것으로 결정하는 경우, 서버는 임시 프로파일을 데이터베이스의 영구 화자 프로파일로 변환하여, 현재 단계(820)에서 화자 프로파일을 업데이트한다. 예를 들어, 서버는 특정 화자로부터 (단계(814)에서와 같이) 임시 성문에 대한 높은 임계값을 충족하는 임계 수 임베딩들을 수신할 때 임시 프로파일이 성숙하다는 결정을 수신한다.

[0275] 단계(818)에서, 서버는 서버가 특정 임베딩에 대한 최대 유사도 점수(809)가 데이터베이스에서 가장 근접하게 일치하는 성문에 대한 더 높은 임계값을 충족하지 못하지만(단계(814)에서), 인바운드 임베딩이 이미 낮은 임계값을 충족되었다고 결정(단계(812)에서)할 때, 약한 임베딩 목록을 업데이트한다. 약한 임베딩 목록은 해당하는 가장 가까운 성문들과 잠재적으로 일치하는 임베딩들을 포함하는 버퍼 또는 격리로 효과적으로 기능한다. 서버는 재클러스터링 동작들과 같은 추후 동작들에서 이러한 약한 임베딩들을 참조하여 가장 가까운 성문의 화자 프로파일에 약한 임베딩들을 포함할지 여부를 결정할 수 있다.

[0276] 단계(822)에서, 서버는 임베딩들 및 화자 정보를 포함하도록 데이터베이스를 업데이트한다. 특정 화자 임베딩에 대해, 데이터베이스는 (단계(816)로부터의) 업데이트된 강한 임베딩 목록, (단계(818)로부터의) 업데이트된 약한 임베딩 목록, 또는 (단계(820)로부터의) 새 데이터 레코드와 같은 하나 이상의 업데이트들을 수신한다. 데이

터베이스는 임베딩, 화자 프로파일 정보, 또는 임베딩 목록과 연관된 화자 식별자(823)로 단계(822)의 업데이트들을 저장한다. 화자 식별자(823)는 콘텐트 서버가 분석 서버에 화자에 대한 임의의 개인 정보를 노출하지 않도록, 콘텐트 시스템의 사용자 식별자를 나타내는 익명화된 값이다.

[0277] 단계(824)에서, 서버는 화자 식별자(823) 및 다운스트림 동작들을 위해 콘텐트 서버에 의해 요청된 화자 프로파일에 관한 임의의 관련 정보를 출력한다. 서버는 사용자 식별자(823)를 예를 들어 미디어 콘텐트 시스템의 컴퓨팅 디바이스들, 최종 사용자 디바이스들, 또는 특정 다운스트림 동작을 수행하는 임의의 다른 디바이스들에 전송한다. 일부 구현들에서, 서버는 또한 데이터베이스 레코드에 저장된 화자 프로파일로부터의 추가 화자 프로파일 정보를 전송한다.

[0278] **혼합된 능동적-수동적 및 지속적 등록**

[0279] 도 9는 혼합된 능동적-수동적 및 지속적 등록 구성들을 사용하여 오디오 신호들을 오디오 처리하기 위한 방법(900)의 동작 단계들을 도시한다. 서버, 또는 다른 컴퓨팅 디바이스는 방법(900)을 실행하지만, 서버는 일부 실시예들에서 등록 단계 동안 방법(900)을 실행할 수 있다. 서버는 능동적 등록 단계 및 배포 단계 동안 방법(900)을 실행한다. 능동적 등록 단계 동안, 서버는 하나 이상의 오디오 프롬프트들 및/또는 시각적 프롬프트들에 응답하는 등록된 화자로부터 등록 오디오 신호들을 수신한다. 프롬프트들은 등록된 화자에게 청각적으로 응답하도록 요청하며, 서버는 등록된 성문들을 생성하고, 특정 화자를 등록하고, 데이터베이스에 해당 화자 프로파일 데이터를 생성하기 위해 등록 오디오 신호들로 수신한다. 배포 단계 동안, 서버는 서버가 특정 인바운드 화자를 식별하기 위해 평가하는 인바운드 오디오 신호들을 수신한다. 서버는 또한 알 수 없는 화자들을 지속적으로 수동적으로 등록할 수 있다. 서버가 인바운드 화자를 식별할 수 없는 경우, 서버는 알 수 없는 화자를 수동으로 등록하고 데이터베이스의 새 화자에 대한 새 데이터 레코드를 생성한다.

[0280] 단계(902)에서, 서버는 화자-사용자의 발화들을 포함하는 등록된 사용자로부터 등록 오디오 신호를 수신한다. 서버는 중개 콘텐트 서버를 통해 최종 사용자 디바이스로부터 등록 오디오 신호들을 수신한다. 서버는 등록 오디오 신호들로부터 특징들을 추출한다. 특징들은 스펙트럼-시간적 특징들일 수 있으며, 일부 경우에, 다양한 유형의 데이터 또는 메타데이터일 수 있다. 서버는 또한 다양한 데이터 증강 동작들과 같은 등록 오디오 신호에 대한 다양한 전처리 동작들을 수행할 수 있다. 서버는 멜 주파수 켑스트럼 계수들(MFCC들), 멜 필터 뱅크들, 선형 필터 뱅크들, 병목 특징들 등을 포함하는 발화들로부터 특징들을 추출할 수 있다. 일부 경우에, 서버는 특징들을 추출하고 화자 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 등록 오디오 신호로부터 특징들을 추출한다.

[0281] 단계(904)에서, 서버는 입력 오디오 신호로부터 추출된 특징들에 대해 임의 개수의 기계 학습 모델들을 포함하는 기계 학습 아키텍처를 적용한다. 기계 학습 아키텍처의 임베딩 추출 모델은 입력 오디오 신호로부터 추출된 특징들을 사용하여 인바운드 화자에 대한 인바운드 임베딩을 추출한다. 서버는 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기)을 포함하는 기계 학습 아키텍처를 적용하여 등록 신호로부터 임베딩들을 추출한다.

[0282] 등록 신호의 유형은 서버가 임베딩들을 추출하는 방법을 지시할 수 있다. 예를 들어, 사용자는 지문을 등록 신호로 제공할 수 있다. 다른 예들에서, 서버는 능선, 골, 세부 사항을 포함하여 디지털 지문 이미지와 관련된 특징들을 추출할 수 있고; 및/또는 서버는 이미지 데이터의 특징들을 추출하도록 구성된 기계 학습 모델들을 사용하여 특징들을 추출할 수 있다. 서버는 화자 임베딩들과 함께, 사용자 인식에 사용되는 특정 유형의 생체 인식에 대해 특징들을 추출하고 해당 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 다양한 유형의 특징들을 추출한다.

[0283] 결정 단계(906)에서, 서버는 등록 성문이 성숙되었는지 여부를 결정한다. 서버는 서버가 특정 신호의 사용자를 수학적으로 식별할 수 있도록 임베딩들 또는 기타 정보의 임계 수를 충족하기에 충분한 임베딩들(또는 생체 인식 정보)을 추출한 경우 등록이 성숙되었다고 결정할 수 있다. 예를 들어, 서버는 서버가 등록자로부터 발화들을 포함하는 하나 이상의 오디오 신호들로부터 순 음성의 임계 지속시간을 수신할 때 등록이 성숙하다고 결정할 수 있다. 추가로 또는 대안으로, 서버가 미리 결정된 수의 다양한 등록 신호들(예를 들어, 2개의 지문 스캔 및 2개의 발화, 5개의 상이한 발화)을 수신할 때 등록이 성숙되었다고 결정할 수 있다.

[0284] 등록이 불완전한(예를 들어, 미성숙한) 경우, 서버는 (단계(902)에서와 같이) 등록 신호에 대해 사용자에게 프롬프트한다. 서버는 등록이 성숙될 때까지 추가 등록 신호들(등록 발화들이 있음)에 대해 사용자에게 프롬프트한다. 일 예로서, 서버가 제1 유형의 콘텐트(예를 들어, 사용자 이름)가 포함된 등록 발화를 수신하는 경우, 서버는 사용자에게 제2 발화(예를 들어, 사용자 생일)을 프롬프트한다. 다른 예로서, 서버가 생체 인식 정보(예를

들어, 지문)를 수신하는 경우, 서버는 등록 발화를 포함하는 오디오 신호를 사용자에게 프롬프트한다.

[0285] 등록이 성숙되는 경우, 서버는 단계(908)로 진행한다. 단계(908)에서, 서버는 등록된 화자-사용자에 대한 새 등록자 성문을 생성한다. 서버는 등록된 화자-사용자에 대한 성문을 추출하기 위해 등록 임베딩들을 통계적으로 또는 알고리즘적으로 결합한다. 서버는 등록자에 대한 임의의 다양한 기타 비식별(non-identifying) 정보와 함께 등록 성문을 화자 프로파일에 저장한다. 서버는 또한 새로운 화자 식별자를 생성하거나 콘텐트 서버에 새로운 화자 식별자를 요청할 수 있다.

[0286] 단계(910)에서, 서버는 예를 들어 새 화자 프로파일에 새로 등록된 성문 및 기타 정보(예를 들어, 화자 식별자, 가입자 식별자)를 저장하여 데이터베이스를 업데이트함으로써, 새로운 화자-사용자를 등록(registering)/등록(enrolling)한다.

[0287] 단계(914)에서, 서버는 새 화자 식별자 및 다운스트림 동작들을 위해 콘텐트 서버에 의해 요청된 새 화자 프로파일에 관한 임의의 관련 정보를 출력한다. 서버는 새로운 화자 식별자를 예를 들어 미디어 콘텐트 시스템의 컴퓨팅 디바이스들, 최종 사용자 디바이스들, 또는 특정 다운스트림 동작을 수행하는 임의의 다른 디바이스들에 전송한다. 일부 구현들에서, 서버는 또한 새로운 데이터베이스 레코드에 저장된 새로운 화자 프로파일로부터의 추가의 새로운 화자 프로파일 정보를 전송한다.

[0288] 능동적 등록 단계 후에, 서버는 등록된 화자들에 대한 인바운드 오디오 신호들을 평가하는 배포 단계로 기계 학습 아키텍처에 들어가지만, 서버는 인식되지 않은 새 화자들을 계속 수동적으로 등록한다.

[0289] 단계(916)에서, 중개 콘텐트 서버를 통해 최종 사용자 디바이스로부터 하나 이상의 인바운드 화자들의 하나 이상의 발화들을 포함하는 인바운드 오디오 신호를 수신한다. 서버는 인바운드 오디오 신호로부터 특징들을 추출한다. 특징들은 스펙트럼-시간적 특징들일 수 있으며, 일부 경우에, 다양한 유형의 데이터 또는 메타데이터일 수 있다. 서버는 또한 다양한 데이터 증강 동작들과 같은 인바운드 오디오 신호에 대한 다양한 전처리 동작들을 수행할 수 있다. 서버는 멜 주파수 켑스트럼 계수들(MFCC들), 멜 필터 뱅크들, 선형 필터 뱅크들, 병목 특징들 등을 포함하는 발화들로부터 특징들을 추출할 수 있다. 일부 경우에, 서버는 특징들을 추출하고 화자 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 인바운드 오디오 신호로부터 특징들을 추출한다. 일부 구성들에서, 서버는 인바운드 신호에 대한 전처리 동작들을 수행(예를 들어, 인바운드 신호 분할, 인바운드 신호 스케일링, 인바운드 신호 노이즈 제거)할 수 있다.

[0290] 단계(918)에서, 서버는 인바운드 오디오 신호에 대한 기계 학습 아키텍처를 적용하여 인바운드 화자들 각각에 대한 임베딩을 추출한다. 서버는 인바운드 오디오 신호로부터 추출된 특징들을 기반으로 인바운드 오디오 신호로부터 각 인바운드 임베딩들을 추출한다.

[0291] 단계(920)에서, 서버는 특정 인바운드 임베딩과 데이터베이스에 저장된 각 성문 사이의 상대적 거리에 기초하여 발화들에 대해 생성된 각 인바운드 화자 임베딩에 대한 유사도 점수를 생성한다. 각 특정 화자에 대해, 서버는 특정 성문에 (유사도 점수에 의해) 가장 근접하게 일치하는 특정 인바운드 화자를 나타내는 최대 유사도 점수를 출력한다.

[0292] 결정 단계(922)에서, 각 특정 인바운드 화자 임베딩에 대해, 서버는 대응하는 최대 유사도 점수가 낮은 유사도 임계값을 충족하는지 여부를 결정한다. 일부 경우에, 낮은 유사도 임계값은 미리 구성된 기본값 또는 특정 성문 및 추정 알려진 화자에 맞게 조정된 적응 임계값이다(도 4에서와 같이).

[0293] 단계(924)에서, 서버는 서버가 특정 인바운드 임베딩에 대한 최대 유사도 점수가 낮은 유사도 임계값을 충족하지 못한다고 결정(단계(922)에서)할 때, 데이터베이스에 새로운 화자 프로파일을 생성한다. 서버는 새 화자 식별자를 새 화자 프로파일에 할당하며, 여기서 서버는 제3자 서버로부터 화자 식별자를 수신한다. 일부 구현들에서, 서버가 제3자 서버에 의해 사용되는 해당 화자 식별자의 해시된(또는 난독화된) 버전을 수신하고, 이로써 서버가 특정 화자에 대한 임의의 개인 식별 정보를 수신하지 못하도록 하여 화자 프라이버시를 유지한다.

[0294] 결정 단계(926)에서, 서버는 서버가 최대 유사도 점수가 낮은 임계값을 충족한다고 결정(단계(922)에서)하는 경우, 최대 유사도 점수가 높은 유사도 임계값을 충족하는지 여부를 결정한다. 일부 경우에, 높은 유사도 임계값은 미리 구성된 기본값 또는 특정 성문 및 추정 등록된 화자에 맞게 조정된 적응 임계값이다(도 4에서와 같이).

[0295] 단계(928)에서, 서버는 서버가 특정 임베딩에 대한 최대 유사도 점수가 데이터베이스에서 가장 근접하게 일치하는 성문에 대한 더 높은 임계값을 충족하지 못하지만(단계(926)에서), 인바운드 임베딩이 이미 낮은 임계값을 충족되었다고 결정(단계(922)에서)할 경우, 약한 임베딩 목록을 업데이트한다. 약한 임베딩 목록은 해당하는 가

장 가까운 성문들과 잠재적으로 일치하는 임베딩들을 포함하는 버퍼 또는 격리로 효과적으로 기능한다. 서버는 재클러스터링 동작들과 같은 추후 동작들에서 이러한 약한 임베딩들을 참조하여 가장 가까운 성문의 화자 프로파일에 약한 임베딩들을 포함할지 여부를 결정할 수 있다.

[0296] 단계(930)에서, 서버가 최대 유사도 점수가 높은 임계값을 충족한다고 결정(단계(814)에서)할 때, 강한 임베딩 목록을 업데이트한다. 서버는 성문들을 생성하기 위해 강한 임베딩들을 사용한다. 예를 들어, 서버는 특정 인바운드 강한 임베딩을 사용하여 특정 성문을 업데이트한다. 서버는 업데이트된 성문을 포함하도록 해당 화자 프로파일을 추가로 업데이트할 수 있다. 특정 임베딩에 대한 최대 유사도 점수가 높은 임계값을 충족했기 때문에, 서버는 특정 인바운드 임베딩이 추정 등록된 화자일 가능성이 있다고 결정한다.

[0297] 단계(910)에서, 서버는 임베딩들 및 화자 정보를 포함하도록 데이터베이스를 업데이트한다. 특정 화자 임베딩에 대해, 데이터베이스는 (단계(930)로부터의) 업데이트된 강한 임베딩 목록, (단계(928)로부터의) 업데이트된 약한 임베딩 목록, 또는 (단계(924)로부터의) 새 데이터 레코드와 같은 하나 이상의 업데이트들을 수신한다. 데이터베이스는 임베딩, 화자 프로파일 정보, 또는 임베딩 목록과 연관된 화자 식별자로 단계(910)의 업데이트들을 저장한다. 화자 식별자는 콘텐트 서버가 분석 서버에 화자에 대한 임의의 개인 정보를 노출하지 않도록, 콘텐트 시스템의 사용자 식별자를 나타내는 익명화된 값이다.

[0298] 단계(914)에서, 서버는 화자 식별자들, 새로운 또는 알려진 식별자들 및 다운스트림 동작들을 위해 콘텐트 서버에 의해 요청된 화자 프로파일에 관한 임의의 관련 정보를 출력한다. 서버는 사용자 식별자를 예를 들어 미디어 콘텐트 시스템의 컴퓨팅 디바이스들, 최종 사용자 디바이스들, 또는 특정 다운스트림 동작을 수행하는 임의의 다른 디바이스들에 전송한다. 일부 구현들에서, 서버는 또한 데이터베이스 레코드에 저장된 화자 프로파일로부터의 추가 화자 프로파일 정보를 전송한다.

[0299] **능동적 및 지속적 등록**

[0300] 도 10은 능동적 및 지속적 등록 구성들을 사용하여 오디오 신호들을 오디오 처리하기 위한 방법(1000)의 동작 단계들을 도시한다. 서버, 또는 다른 컴퓨팅 디바이스는 방법(900)을 실행하지만, 서버는 일부 실시예들에서 등록 단계 동안 방법(900)을 실행할 수 있다. 서버는 능동적 등록 단계 및 배포 단계 동안 방법(900)을 실행한다. 능동적 등록 단계 동안, 서버는 하나 이상의 오디오 프롬프트들 및/또는 시각적 프롬프트들에 응답하는 등록된 화자로부터 등록 오디오 신호들을 수신한다. 프롬프트들은 등록된 화자에게 청각적으로 응답하도록 요청하며, 서버는 등록된 성문들을 생성하고, 특정 화자를 등록하고, 데이터베이스에 해당 화자 프로파일 데이터를 생성하기 위해 등록 오디오 신호들로 수신한다. 배포 단계 동안, 서버는 서버가 특정 인바운드 화자를 식별하기 위해 평가하는 인바운드 오디오 신호들을 수신한다. 도 9의 실시예와 달리, 능동적 등록 단계는 모든 사용자들을 등록하는 데 필수이다. 서버가 오디오 신호들을 지속적으로, 수동적으로 평가하더라도, 서버는 인식할 수 없는 새 화자들을 수동으로 등록하지 않는다. 클러스터링 동작들은 준-지도(semi-supervised)되지만, 서버는 미리 알려진 개수의 클러스터들로 구성되어 있기 때문에 클러스터링은 다른 실시예들보다 더 제한되거나 더 단순화될 수 있다. 예를 들어, 데이터베이스는 등록을 완료한 미리 구성된 명수의 등록된 화자들을 할당한다.

[0301] 단계(1002)에서, 서버는 화자-사용자의 발화들을 포함하는 등록된 사용자로부터 등록 오디오 신호를 수신한다. 서버는 중개 콘텐트 서버를 통해 최종 사용자 디바이스로부터 등록 오디오 신호들을 수신한다. 서버는 등록 오디오 신호들로부터 특징들을 추출한다. 특징들은 스펙트럼-시간적 특징들일 수 있으며, 일부 경우에, 다양한 유형의 데이터 또는 메타데이터일 수 있다. 서버는 또한 다양한 데이터 증강 동작들과 같은 등록 오디오 신호에 대한 다양한 전처리 동작들을 수행할 수 있다. 서버는 멜 주파수 켑스트럼 계수들(MFCC들), 멜 필터 뱅크들, 선형 필터 뱅크들, 병목 특징들 등을 포함하는 발화들로부터 특징들을 추출할 수 있다. 일부 경우에, 서버는 특징들을 추출하고 화자 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 등록 오디오 신호로부터 특징들을 추출한다.

[0302] 단계(1004)에서, 서버는 입력 오디오 신호로부터 추출된 특징들에 대해 임의 개수의 기계 학습 모델들을 포함하는 기계 학습 아키텍처를 적용한다. 기계 학습 아키텍처의 임베딩 추출 모델은 입력 오디오 신호로부터 추출된 특징들을 사용하여 인바운드 화자에 대한 인바운드 임베딩을 추출한다. 서버는 다양한 기계 학습 모델들(예를 들어, 임베딩 추출기)을 포함하는 기계 학습 아키텍처를 적용하여 등록 신호로부터 임베딩들을 추출한다.

[0303] 결정 단계(1006)에서, 서버는 등록 성문이 성숙되었는지 여부를 결정한다. 서버는 서버가 특정 신호의 사용자를 수학적으로 식별할 수 있도록 임베딩들 또는 기타 정보의 임계 수를 충족하기에 충분한 임베딩들(또는 생체 인식 정보)을 추출한 경우 등록이 성숙되었다고 결정할 수 있다. 예를 들어, 서버는 서버가 등록자로부터 발화들

을 포함하는 하나 이상의 오디오 신호들로부터 순 음성의 임계 지속시간을 수신할 때 등록이 성숙하다고 결정할 수 있다. 추가로 또는 대안으로, 서버가 미리 결정된 수의 다양한 등록 신호들(예를 들어, 2개의 지문 스캔 및 2개의 발화, 5개의 상이한 발화)을 수신할 때 등록이 성숙되었다고 결정할 수 있다.

[0304] 등록이 불완전한(예를 들어, 미성숙한) 경우, 서버는 (단계(1006)에서와 같이) 등록 신호에 대해 사용자에게 프롬프트한다. 서버는 등록이 성숙될 때까지 추가 등록 신호들(등록 발화들이 있음)에 대해 사용자에게 프롬프트한다. 일 예로서, 서버가 제1 유형의 콘텐트(예를 들어, 사용자 이름)가 포함된 등록 발화를 수신하는 경우, 서버는 사용자에게 제2 발화(예를 들어, 사용자 생일)을 프롬프트한다. 다른 예로서, 서버가 생체 인식 정보(예를 들어, 지문)를 수신하는 경우, 서버는 등록 발화를 포함하는 오디오 신호를 사용자에게 프롬프트한다.

[0305] 등록이 성숙되는 경우, 서버는 단계(1008)로 진행한다. 단계(1008)에서, 서버는 등록된 화자-사용자에 대한 새 등록자 성문을 생성한다. 서버는 등록된 화자-사용자에 대한 성문을 추출하기 위해 등록 임베딩들을 통계적으로 또는 알고리즘적으로 결합한다. 서버는 등록자에 대한 임의의 다양한 기타 비식별 정보와 함께 등록 성문을 화자 프로파일에 저장한다. 서버는 또한 새로운 화자 식별자를 생성하거나 콘텐트 서버에 새로운 화자 식별자를 요청할 수 있다.

[0306] 단계(1010)에서, 서버는 예를 들어 새 화자 프로파일에 새로 등록된 성문 및 기타 정보(예를 들어, 화자 식별자, 가입자 식별자)를 저장하여 데이터베이스를 업데이트함으로써, 새로운 화자-사용자를 등록(registering)/등록(enrolling)한다.

[0307] 단계(1014)에서, 서버는 새 화자 식별자 및 다운스트림 동작들을 위해 콘텐트 서버에 의해 요청된 새 화자 프로파일에 관한 임의의 관련 정보를 출력한다. 서버는 새로운 화자 식별자를 예를 들어 미디어 콘텐트 시스템의 컴퓨팅 디바이스들, 최종 사용자 디바이스들, 또는 특정 다운스트림 동작을 수행하는 임의의 다른 디바이스들에 전송한다. 일부 구현들에서, 서버는 또한 새로운 데이터베이스 레코드에 저장된 새로운 화자 프로파일로부터의 추가의 새로운 화자 프로파일 정보를 전송한다.

[0308] 능동적 등록 단계 후에, 서버는 등록된 화자들에 대한 인바운드 오디오 신호들을 평가하는 배포 단계로 기계 학습 아키텍처에 들어가지만, 서버는 인식되지 않은 새 화자들을 계속 수동적으로 등록한다.

[0309] 단계(1016)에서, 중개 콘텐트 서버를 통해 최종 사용자 디바이스로부터 하나 이상의 인바운드 화자들의 하나 이상의 발화들을 포함하는 인바운드 오디오 신호를 수신한다. 서버는 인바운드 오디오 신호로부터 특징들을 추출한다. 특징들은 스펙트럼-시간적 특징들일 수 있으며, 일부 경우에, 다양한 유형의 데이터 또는 메타데이터일 수 있다. 서버는 또한 다양한 데이터 증강 동작들과 같은 인바운드 오디오 신호에 대한 다양한 전처리 동작들을 수행할 수 있다. 서버는 멜 주파수 켑스트럼 계수들(MFCC들), 멜 필터 뱅크들, 선형 필터 뱅크들, 병목 특징들 등을 포함하는 발화들로부터 특징들을 추출할 수 있다. 일부 경우에, 서버는 특징들을 추출하고 화자 임베딩들을 생성하도록 구성된 기계 학습 모델들을 사용하여 인바운드 오디오 신호로부터 특징들을 추출한다. 일부 구성들에서, 서버는 인바운드 신호에 대한 전처리 동작들을 수행(예를 들어, 인바운드 신호 분할, 인바운드 신호 스케일링, 인바운드 신호 노이즈 제거)할 수 있다.

[0310] 단계(1018)에서, 서버는 인바운드 오디오 신호에 대한 기계 학습 아키텍처를 적용하여 인바운드 화자들 각각에 대한 임베딩을 추출한다. 서버는 인바운드 오디오 신호로부터 추출된 특징들을 기반으로 인바운드 오디오 신호로부터 각 인바운드 임베딩들을 추출한다.

[0311] 단계(1020)에서, 서버는 특정 인바운드 임베딩과 데이터베이스에 저장된 각 성문 사이의 상대적 거리에 기초하여 발화들에 대해 생성된 각 인바운드 화자 임베딩에 대한 유사도 점수를 생성한다. 각 특정 화자에 대해, 서버는 특정 성문에 (유사도 점수에 의해) 가장 근접하게 일치하는 특정 인바운드 화자를 나타내는 최대 유사도 점수를 출력한다.

[0312] 결정 단계(1022)에서, 각 특정 인바운드 화자 임베딩에 대해, 서버는 대응하는 최대 유사도 점수가 낮은 유사도 임계값을 충족하는지 여부를 결정한다. 일부 경우에, 낮은 유사도 임계값은 미리 구성된 기본값 또는 특정 성문 및 추정 알려진 화자에 맞게 조정된 적응 임계값이다(도 4에서와 같이).

[0313] 결정 단계(1026)에서, 서버는 서버가 최대 유사도 점수가 낮은 임계값을 충족한다고 결정(단계(1022)에서)하는 경우, 최대 유사도 점수가 높은 유사도 임계값을 충족하는지 여부를 결정한다. 일부 경우에, 높은 유사도 임계값은 미리 구성된 기본값 또는 특정 성문 및 추정 등록된 화자에 맞게 조정된 적응 임계값이다(도 4에서와 같이).

[0314] 단계(1028)에서, 서버는 서버가 특정 임베딩에 대한 최대 유사도 점수가 데이터베이스에서 가장 근접하게 일치하는 성문에 대한 더 높은 임계값을 충족하지 못하지만(단계(1026)에서), 인바운드 임베딩이 이미 낮은 임계값을 충족되었다고 결정(단계(1022)에서)할 경우, 약한 임베딩 목록을 업데이트한다. 약한 임베딩 목록은 해당하는 가장 가까운 성문들과 잠재적으로 일치하는 임베딩들을 포함하는 버퍼 또는 격리로 효과적으로 기능한다. 서버는 재클러스터링 동작들과 같은 추후 동작들에서 이러한 약한 임베딩들을 참조하여 가장 가까운 성문의 화자 프로파일에 약한 임베딩들을 포함할지 여부를 결정할 수 있다.

[0315] 단계(1026)에서, 서버가 최대 유사도 점수가 높은 임계값을 충족한다고 결정(단계(1026)에서)할 때, 강한 임베딩 목록을 업데이트한다. 서버는 성문들을 생성하기 위해 강한 임베딩들을 사용한다. 예를 들어, 서버는 특정 인바운드 강한 임베딩을 사용하여 특정 성문을 업데이트한다. 서버는 업데이트된 성문을 포함하도록 해당 화자 프로파일을 추가로 업데이트할 수 있다. 특정 임베딩에 대한 최대 유사도 점수가 높은 임계값을 충족했기 때문에, 서버는 특정 인바운드 임베딩이 추정 등록된 화자일 가능성이 있다고 결정한다.

[0316] 단계(1010)에서, 서버는 임베딩들 및 화자 정보를 포함하도록 데이터베이스를 업데이트한다. 특정 화자 임베딩에 대해, 데이터베이스는 (단계(1030)로부터의) 업데이트된 강한 임베딩 목록 또는 (단계(1028)로부터의) 업데이트된 약한 임베딩 목록과 같은 하나 이상의 업데이트들을 수신한다. 데이터베이스는 임베딩, 화자 프로파일 정보, 또는 임베딩 목록과 연관된 화자 식별자로 단계(1010)의 업데이트들을 저장한다. 화자 식별자는 콘텐트 서버가 분석 서버에 화자에 대한 임의의 개인 정보를 노출하지 않도록, 콘텐트 시스템의 사용자 식별자를 나타내는 익명화된 값이다.

[0317] 단계(1024)에서, 서버는 서버가 특정 인바운드 화자 임베딩이 낮은 임계값에 못미친다고 결정(단계(1022)에서)하는 경우, 중개 콘텐트 서버 또는 최종 사용자 디바이스에 대해 화자가 인식되지 않음을 나타내는 경고 또는 다른 명령어를 생성한다. 서버는 콘텐트 서버에 추가 정보를 제공하도록 미리 구성될 수도 있다.

[0318] 단계(1014)에서, 서버는 콘텐트 서버에 다양한 유형의 정보를 출력한다. 일부 경우에, 서버는 화자 식별자들 및 다운스트림 동작들을 위해 콘텐트 서버에 의해 요청된 화자 프로파일에 관한 임의의 관련 정보를 전송한다. 이러한 경우에, 서버는 사용자 식별자를 예를 들어 미디어 콘텐트 시스템의 컴퓨팅 디바이스들, 최종 사용자 디바이스들, 또는 특정 다운스트림 동작을 수행하는 임의의 다른 디바이스들에 전송한다. 일부 구현들에서, 서버는 또한 데이터베이스 레코드에 저장된 화자 프로파일로부터의 추가 화자 프로파일 정보를 전송한다. 대안적으로, 서버는 특정 화자가 서버에 의해 인식되지 않았음을 나타내는 경고 또는 기타 명령어들을 콘텐트 서버에 전송한다.

[0319] **추가적인 예시적인 실시예들**

[0320] 도 11a 내지 11b는 기계 학습 모델들이 차량 또는 다른 에지 디바이스(예를 들어, 자동차, 홈 어시스턴트 디바이스, 스마트 기기)에 의해 구현되는 오디오 처리 기계 학습 동작들을 사용하는 시스템(1100)의 컴포넌트들을 도시한다.

[0321] 차량은 음성을 포함하는 오디오 파동들(1110)을 캡처하고 오디오 파동들(1110)을 오디오 처리 동작들을 위한 오디오 신호들로 변환하도록 구성된 마이크(1108)를 포함한다. 차량은 본원에 설명된 다양한 오디오 처리 동작들을 수행하도록 구성된 컴퓨팅 하드웨어 및 소프트웨어 컴포넌트들(분석 컴퓨터(1102) 및 화자 데이터베이스(1104)로 표시됨)를 포함한다. 시스템(1100)에 설명된 컴포넌트들 및 동작들은 도 1 내지 2의 컴포넌트들 및 동작들과 유사하다. 도 1의 시스템(100)은 분석 서버(102)에 기계 학습 오디오 처리 동작들의 대부분을 배치했지만, 콘텐트 시스템(110)이 일부 실시예들에서 특정 동작들을 수행할 수 있다. 도 2의 시스템(200)은 최종 사용자 디바이스(214)에 기계 학습 오디오 처리 동작들의 대부분을 배치했지만, 최종 사용자 디바이스(214)는 다양한 동작들 및 데이터베이스 정보에 대해 분석 시스템(201) 또는 콘텐트 시스템(210)에 여전히 의존할 수 있다. 그러나 도 11의 차량 기반 시스템은 차량 기반 시스템(1100) 내에서 오디오 처리 동작들 및 데이터의 많은 부분을 캡슐화하려고 하며, 외부 시스템 인프라스트럭처의 디바이스들에 상대적으로 덜 의존한다.

[0322] 분석 컴퓨터(1102)는 마이크(1108)로부터 입력 데이터 신호들을 수신하고 발화들을 식별하기 위해 VAD 및 ASR과 같은 다양한 전처리 동작들을 수행한다. 분석 컴퓨터(1102)는 임의 개수의 기계 학습 모델들을 적용하여 특징들을 추출하고, 임베딩들을 추출하고, 임베딩들을 화자 데이터베이스(1104)에 저장된 성문들과 비교한다. 분석 컴퓨터(1102)는 인포테인먼트 시스템, 엔진, 도어록, 및 차량의 다른 컴포넌트들과 같은 차량의 다양한 전자 컴포넌트들에 결합된다. 분석 컴퓨터(1102)는 차량의 다양한 옵션들을 활성화하거나 조정하기 위해 운전자 또는 승객들로부터 음성 명령어들을 수신한다.

[0323] 일부 실시예들에서, 분석 컴퓨터(1102)는 차량의 기능들에 대한 부모 제어 동작들 또는 기타 제한들을 사용한다. 화자 데이터베이스(1104)에 저장된 화자 프로파일은 분석 컴퓨터(1102)가 특정 동작들을 수행하는 것을 금지하는 특정 기능적 제한들을 포함한다. 예를 들어, 분석 컴퓨터(1102)는 임베딩 추출 모델을 실행함으로써, 또는 대안적으로 음성을 사용하여 연령을 결정하기 위한 알려진 기계 학습 모델을 실행함으로써 아동 화자 프로파일에 등록된 화자 임베딩들을 검출할 수 있다. 그런 다음 분석 컴퓨터(1102)는 스타터(starter) 또는 이그니션(ignition)의 활성화를 금지함으로써, 분석 컴퓨터(1102)가 화자 데이터베이스(1104)에서 인가된 사용자의 성문과 일치하는 임베딩을 검출할 때까지 엔진이 활성화되는 것을 방지한다. 이 기능은 부모 통제일 뿐만 아니라 도난 방지에 적합하며, 이로써 분석 컴퓨터(1102)는 분석 컴퓨터(1102)가 등록된 화자-사용자로부터의 인바운드 임베딩을 포함하는 인바운드 오디오 신호를 긍정적으로 감지할 때까지 엔진이 활성화되는 것을 방지한다.

[0324] 분석 컴퓨터(1102)는 드라이버(예를 들어, 제1 부모), 이차 드라이버(예를 들어, 제2 드라이버) 및 승객들(예를 들어, 아동)을 능동적으로 또는 수동적으로 등록한다. 일 예로서, 드라이버가 차량을 처음 구매할 때, 인포테인먼트 디바이스를 통해 디스플레이되는 GUI는 드라이버에게 특정 문구들을 말하도록 요청하는 프롬프트를 표시함으로써, 마이크(1108)에 의해 캡처된 등록 발화들을 제출한다. 그런 다음 분석 컴퓨터(1102)는 본원에 설명된 다양한 프로세스들을 수행하여 운전자에 대한 성문을 생성하고, 분석 컴퓨터(1102)는 이를 화자 데이터베이스(1104)의 운전자에 대한 화자 프로파일에 저장한다. 운전자는 또한 예를 들어 좌석 위치, 라디오 방송국, 어린이 보호용 잠금 장치(child-locks) 활성화, 보안 기본 설정, 헤드라이트 지연 등과 관련된 이름 및 특정 기본 설정과 같은 운전자에 대한 정보를 GUI를 통해 입력한다. 이 화자 정보는 운전자의 화자 프로파일에 저장된다. 운전자는 또한 GUI를 통해 시스템(1100)을 동작시킬 예상되는 다수의 화자들을 입력(예를 들어, 차량에 탑승)할 수 있다. 분석 컴퓨터(1102)는 예상된 화자들의 수에 따라 클러스터들 및 성문들을 생성한다. 분석 컴퓨터(1102)는 예상된 화자들의 수에 따라 화자 프로파일을 생성할 수 있거나 분석 컴퓨터(1102)는 본원에 설명된 능동적 등록, 수동적 등록, 및/또는 지속적 등록 동작들을 실행하여 화자 프로파일들을 생성한다. 예를 들어, 분석 컴퓨터(1102)는 아동 화자 프로파일에 대한 성문을 생성하기 위해 지속적이고 수동적인 등록 동작들을 수행한다. 운전자 또는 아동은 인포테인먼트 시스템의 GUI를 통해 아동에 대한 다양한 유형의 화자 정보를 입력할 수 있다. 이 화자 정보에는 예를 들어 본원에 언급된 부모 제어들과 같은 운전자 권한이 포함될 수 있다.

[0325] 일부 실시예들에서, 분석 컴퓨터(1102)는 정적 등록 구성을 사용함으로써, 분석 컴퓨터(1102)는 알 수 없는 화자 임베딩들을 새로운 등록들로 수락하지 않는다. 추가로, 분석 컴퓨터(1102)는 인식되지 않은 성문들의 인증을 거부하고 화자가 차량의 특정 기능들에 액세스하는 것을 허용하지 않는 인증 기능들을 수행한다. 예를 들어, 분석 컴퓨터(1102)는 차량 및 차량 동작에 대한 무단 액세스를 제한하기 위해 차량들(예를 들어, 경찰차, 배달 트럭)에 사용될 수 있다.

[0326] 분석 컴퓨터(1102)는 분석 컴퓨터(1102)가 화자 성문들뿐만 아니라 화자 정보를 나타내는 다양한 유형의 데이터에 따라 분석 데이터베이스(1104)에 저장하는 화자 프로파일들을 생성하기 위한 지속적 등록 동작들을 수행한다. 화자 정보의 비제한적인 예들에는 시동을 켜거나 활성화하는 것, 문(예를 들어, 어린이 보호용 잠금 장치)을 여는 것 등과 같은 차량의 다양한 컴퓨터 기반 특징들에 대한 동작 권한들이 포함될 수 있다. 일 예로서, 인포테인먼트 시스템의 GUI를 통한 운전자(예를 들어, 부모) 입력들은 화자 프로파일의 화자 식별자 및 아동의 성문과 연관된 아동 화자 프로파일 및 아동 이름의 다양한 권한들을 나타낸다. 분석 컴퓨터(1102)가 새로운 화자(아동)에 대한 새로운 성문을 식별할 때, 인포테인먼트 시스템은 분석 컴퓨터(1102)가 아동의 새로운 성문을 식별 및 생성했음을 운전자(또는 다른 알려진 등록된 화자)(예를 들어, 부모)를 나타내는 GUI 프롬프트를 생성한다. 부모는 특정 화자 프로파일(또는 화자 식별자)이 아동에 대한 새 성문과 연관되어 있는지 확인하는 하나 이상의 입력들을 입력할 수 있다. 그런 다음 분석 컴퓨터(1102)는 새로운 성문을 분석 데이터베이스(1104)의 아동의 화자 프로파일에 저장한다. 아동은 GUI를 통해 좌석 선호도 또는 기후 선호도에 대한 선호도 구성들과 같은 다양한 유형의 화자 정보를 화자 프로파일에 입력할 수 있다. 아동을 등록한 후, 분석 컴퓨터(1102)는 마이크(1108)에 의해 수신된 발화들에 따라 아동이 차량에 있을 때 수동적으로 식별한다. 그런 다음 분석 컴퓨터(1102)는 인포테인먼트 시스템 또는 관련 제어 시스템들이 아동의 화자 프로파일의 선호도 데이터에 따라 기능(예를 들어, 좌석 위치를 변경)하도록 지시한다.

[0327] 일 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하는 단계; 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 아키텍처를 적용하여 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들을 추출하는 단계; 복수의 인바운드 화자들의 각 인바운드 화자에 대해, 컴퓨터에 의해, 인바운드 화자에 대한 인바운드 임베딩에 기초하여 하나 이상의 유사도 점

수들을 생성하는 단계로서, 인바운드 화자에 대한 각 유사도 점수는 화자 프로파일 데이터베이스에 저장된 인바운드 임베딩과 하나 이상의 성문들 사이의 거리를 나타내는, 상기 생성하는 단계; 및 컴퓨터에 의해, 하나 이상의 성문들로부터 인바운드 화자에 대한 가장 가까운 음성을 식별하는 단계로서, 가장 가까운 성문은 인바운드 화자에 대해 생성된 하나 이상의 유사도 점수들 중 최대 유사도 점수에 대응하는, 상기 식별하는 단계; 및 각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족하는 경우, 컴퓨터에 의해, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하는 단계를 포함한다.

[0328] 방법은, 컴퓨터에 의해, 인바운드 오디오 신호와 함께 수신된 가입자 식별자에 기초하여 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들을 식별하는 단계를 더 포함할 수 있다.

[0329] 방법은, 컴퓨터에 의해, 인바운드 화자 임베딩에 대한 최대 유사도 점수가 하나 이상의 유사도 점수들을 충족한다고 결정하는 단계; 및 컴퓨터에 의해, 인바운드 화자에 대한 가장 가까운 성문을 포함하는 화자 데이터베이스의 화자 프로파일을 식별하는 단계로서, 화자 프로파일은 화자 식별자를 포함하는, 상기 식별하는 단계를 더 포함할 수 있다.

[0330] 방법은, 컴퓨터에 의해, 인바운드 화자에 대한 최대 유사도 점수가 제1 유사도 임계값을 충족하고 제2 유사도 임계값에 못 미친다고 결정하는 단계로서, 제1 유사도 임계값이 제2 유사도 임계값보다 비교적 낮은, 상기 결정하는 단계; 및 컴퓨터에 의해, 인바운드 화자의 인바운드 임베딩을 포함하도록 화자 데이터베이스에 저장된 약한 임베딩 목록을 업데이트하는 단계를 더 포함할 수 있다.

[0331] 방법은, 컴퓨터에 의해, 가입자 식별자와 연관된 화자 데이터베이스의 하나 이상의 화자 프로파일들에 대한 재클러스터링 동작들을 실행하는 단계; 컴퓨터에 의해, 재클러스터링 동작에 기초하여 약한 임베딩 목록의 인바운드 임베딩에 대한 최대 유사도 점수를 업데이트하는 단계; 및 약한 임베딩 목록의 인바운드 임베딩에 대한 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족한다고 결정하는 것에 응답하여, 컴퓨터에 의해, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하는 단계를 더 포함할 수 있다.

[0332] 방법은, 컴퓨터에 의해, 가입자 식별자에 대한 재클러스터링 동작을 수행하기 위한 트리거링 조건을 검출하는 단계; 및 컴퓨터에 의해, 가입자 식별자와 연관된 하나 이상의 화자 프로파일들의 복수의 성문들에 대한 계층적 클러스터링 동작을 실행하는 단계를 더 포함할 수 있다.

[0333] 방법은, 컴퓨터에 의해, 기존 화자 프로파일로부터 새로운 성문 클러스터와 연관된 화자 식별자를 업데이트하는 단계로서, 새로운 성문 클러스터는 가입자 식별자와 연관된 하나 이상의 화자 프로파일들에 계층적 클러스터링 동작을 적용함으로써 생성되는, 상기 업데이트하는 단계를 더 포함할 수 있다.

[0334] 방법은, 컴퓨터에 의해, 중개 서버를 통해 최종 사용자 디바이스로부터 인바운드 오디오 신호를 수신하는 단계; 및 컴퓨터에 의해, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 화자 식별자를 전송하는 단계를 더 포함할 수 있다.

[0335] 방법은 컴퓨터에 의해, 복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하는 단계를 더 포함할 수 있다.

[0336] 방법은, 컴퓨터에 의해, 하나 이상의 특징들에 기계 학습 아키텍처를 적용하여 하나 이상의 오디오 이벤트들을 식별하여 환경 설정을 검출하는 단계; 및 컴퓨터에 의해, 환경 설정의 표시자 및 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 각 화자 식별자를 중개 서버에 전송하는 단계를 더 포함할 수 있다.

[0337] 방법은, 각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들에 못 미치는 경우, 컴퓨터에 의해, 인바운드 임베딩을 포함하는 가입자 식별자와 연관된 새로운 화자 프로파일을 생성하는 단계; 및 컴퓨터에 의해, 새로운 화자 프로파일 및 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하는 단계를 더 포함할 수 있다.

[0338] 일 실시예에서, 시스템은 화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 컴퓨터로서, 복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하고; 인바운드 오디오 신호에 기계 학습 아키텍처를 적용하여 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들을 추출하고; 복수의 인바운드 화자들의 각 인바운드 화자

에 대해, 인바운드 화자에 대한 인바운드 임베딩에 기초하여 하나 이상의 유사도 점수들을 생성하되, 인바운드 화자에 대한 각 유사도 점수는 인바운드 임베딩과 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들 사이의 거리를 나타내고; 하나 이상의 성문들로부터 인바운드 화자에 대한 가장 가까운 성문을 식별하되, 가장 가까운 성문은 인바운드 화자에 대해 생성된 하나 이상의 유사도 점수들 중 최대 유사도 점수에 대응하고; 각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족하는 경우, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함한다.

[0339] 컴퓨터는 인바운드 오디오 신호와 함께 수신된 가입자 식별자에 기초하여 화자 프로파일 데이터베이스에 저장된 하나 이상의 성문들을 식별하도록 구성될 수 있다.

[0340] 컴퓨터는 인바운드 화자 임베딩에 대한 최대 유사도 점수가 하나 이상의 유사도 점수들을 충족한다고 결정하고; 인바운드 화자에 대한 가장 가까운 성문을 포함하는 화자 데이터베이스의 화자 프로파일을 식별하되, 화자 프로파일은 화자 식별자를 포함하도록 구성될 수 있다.

[0341] 컴퓨터는 인바운드 화자에 대한 최대 유사도 점수가 제1 유사도 임계값을 충족하고 제2 유사도 임계값에 못 미친다고 결정하되, 제1 유사도 임계값이 제2 유사도 임계값보다 비교적 낮고; 인바운드 화자의 인바운드 임베딩을 포함하도록 화자 데이터베이스에 저장된 약한 임베딩 목록을 업데이트하도록 구성될 수 있다.

[0342] 컴퓨터는 가입자 식별자와 연관된 화자 데이터베이스의 하나 이상의 화자 프로파일들에 대한 재클러스터링 동작을 실행하고; 재클러스터링 동작에 기초하여 약한 임베딩 목록의 인바운드 임베딩에 대한 최대 유사도 점수를 업데이트하고; 약한 임베딩 목록의 인바운드 임베딩에 대한 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들을 충족한다고 결정하는 것에 응답하여, 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 인바운드 화자에 대한 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하도록 구성될 수 있다.

[0343] 컴퓨터는 가입자 식별자에 대한 재클러스터링 동작을 수행하기 위한 트리거링 조건을 검출하고; 가입자 식별자와 연관된 하나 이상의 화자 프로파일들의 복수의 성문들에 대한 계층적 클러스터링 동작을 실행하도록 구성될 수 있다.

[0344] 컴퓨터는 기존 화자 프로파일로부터 새로운 성문 클러스터와 연관된 화자 식별자를 업데이트하되, 새로운 성문 클러스터는 가입자 식별자와 연관된 하나 이상의 화자 프로파일들에 계층적 클러스터링 동작을 적용함으로써 생성되도록 구성될 수 있다.

[0345] 컴퓨터는 중개 서버를 통해 최종 사용자 디바이스로부터 인바운드 오디오 신호를 수신하고; 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 화자 식별자를 전송하도록 구성될 수 있다.

[0346] 컴퓨터는 복수의 인바운드 화자들의 복수의 발화들을 포함하는 인바운드 오디오 신호를 수신하도록 구성될 수 있다.

[0347] 컴퓨터는 하나 이상의 특징들에 기계 학습 아키텍처를 적용하여 하나 이상의 오디오 이벤트들을 식별하여 환경 설정을 검출하고; 하나 이상의 유사도 점수 임계값들을 충족하는 최대 유사도 점수를 갖는 각 인바운드 화자와 연관된 환경 설정 및 각 화자 식별자의 표시자를 중개 서버에 전송하도록 구성될 수 있다.

[0348] 컴퓨터는, 각 최대 유사도 점수가 하나 이상의 유사도 점수 임계값들에 못 미치는 경우, 인바운드 임베딩을 포함하는 가입자 식별자와 연관된 새로운 화자 프로파일을 생성하고; 인바운드 화자에 대한 새로운 화자 프로파일 및 인바운드 임베딩을 포함하도록 화자 프로파일 데이터베이스를 업데이트하도록 구성될 수 있다.

[0349] 일 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 콘텐트 서버를 통해 최종 사용자 디바이스로부터 인바운드 화자에 대한 인바운드 오디오 신호를 수신하는 단계; 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하는 단계; 컴퓨터에 의해, 인바운드 임베딩과 화자 데이터베이스의 화자 프로파일에 저장된 성문 사이의 거리에 기초하여 인바운드 임베딩에 대한 유사도 점수를 생성하는 단계로서, 유사도 점수는 하나 이상의 유사도 점수 임계값들을 충족하는, 상기 생성하는 단계; 상기 컴퓨터에 의해, 화자 프로파일에서 콘텐트 서버에 대한 하나 이상의 콘텐트 특성들에 대응하는 하나 이상의 화자 특성들을 식별하는 단계; 및 컴퓨터에 의해, 인바운드 화자와 연관된 하나 이상의 화자 특성들을 미디어 콘텐트 서버에 전송하는 단계를 포함한다.

[0350] 방법은, 컴퓨터에 의해, 콘텐트 서버로부터 수신된 가입자 식별자에 기초하여 화자 프로파일을 식별하는 단계를 더 포함할 수 있다.

[0351] 컴퓨터에 의해 수신된 가입자 식별자는 제1 익명 식별자일 수 있고, 가입자 식별자는 하나 이상의 화자들에 대한 하나 이상의 화자 식별자들과 연관될 수 있고, 각 화자 식별자는 각 화자 프로파일에 대응하는 제2 익명 식별자일 수 있다.

[0352] 하나 이상의 화자 특성들은 연령 특성 및 성별 특성 중 적어도 하나를 포함할 수 있다.

[0353] 방법은, 컴퓨터에 의해, 인바운드 오디오 신호에 제2 기계 학습 모델을 적용함으로써 인바운드 화자의 연령 특성 및 성별 특성 중 적어도 하나를 결정하는 단계를 더 포함할 수 있다.

[0354] 방법은, 컴퓨터에 의해, 인바운드 화자에 대한 화자 프로파일에 저장된 연령 특성에 기초하여 인바운드 화자의 연령을 결정하는 단계를 더 포함할 수 있다.

[0355] 방법은, 컴퓨터에 의해, 콘텐트 서버로부터 인바운드 화자에 대한 적어도 하나의 화자 특성을 나타내는 화자 정보를 수신하는 단계; 컴퓨터에 의해, 인바운드 화자에 대한 화자 프로파일에 적어도 하나의 화자 특성을 저장하는 단계; 및 컴퓨터에 의해, 화자 프로파일의 적어도 하나의 화자 특성에 부분적으로 기초하여 화자 프로파일을 식별하는 단계를 더 포함할 수 있다.

[0356] 방법은, 컴퓨터에 의해, 인바운드 오디오 신호와 함께 콘텐트 서버로부터 인바운드 인증 데이터를 수신하는 단계; 및 컴퓨터에 의해, 유사도 점수가 유사도 임계값을 충족하고 인바운드 인증 데이터가 화자 프로파일에 저장된 예상된 인증 데이터를 충족하는 것에 기초하여 인바운드 화자를 인증하는 단계를 더 포함할 수 있다.

[0357] 인증 데이터는 최종 사용자 디바이스, 화자 정보 및 생체 인증 정보와 연관된 최종 사용자 디바이스 정보, 메타데이터 중 적어도 하나를 포함할 수 있다.

[0358] 방법은, 컴퓨터에 의해, 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하는 단계; 및 컴퓨터에 의해, 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 상기 하나 이상의 특징들에 기초하여 스푸핑(spoofing) 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하는 단계를 더 포함할 수 있다.

[0359] 방법은, 컴퓨터에 의해, 미디어 콘텐트 데이터베이스에서 하나 이상의 화자 특성들에 대응하는 하나 이상의 콘텐트 특성들을 갖는 하나 이상의 미디어 콘텐트 파일들을 식별하는 단계를 더 포함할 수 있다.

[0360] 방법은, 컴퓨터에 의해, 화자의 하나 이상의 화자 특성들 중 연령 특성이 연령 제한 콘텐트의 하나 이상의 콘텐트 특성들의 해당 연령 제한 특성을 충족하는 것에 기초하여 콘텐트 데이터베이스에서 연령 제한 콘텐트를 활성화하는 단계를 더 포함할 수 있다.

[0361] 일 실시예에서, 시스템은 복수의 화자 프로파일들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 서버로서, 콘텐트 서버를 통해 최종 사용자 디바이스로부터 인바운드 화자에 대한 인바운드 오디오 신호를 수신하고; 인바운드 오디오 신호에 기계 학습 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하고; 인바운드 임베딩과 화자 데이터베이스의 화자 프로파일에 저장된 성문 사이의 거리에 기초하여 인바운드 임베딩에 대한 유사도 점수를 생성하되, 유사도 점수는 하나 이상의 유사도 점수 임계값들을 충족하고; 화자 프로파일에서 콘텐트 서버에 대한 하나 이상의 콘텐트 특성들에 대응하는 하나 이상의 화자 특성들을 식별하고; 인바운드 화자와 연관된 하나 이상의 화자 특성들을 미디어 콘텐트 서버에 전송하도록 구성된 프로세서를 포함하는, 상기 서버를 포함한다.

[0362] 컴퓨터는 콘텐트 서버로부터 수신된 가입자 식별자에 기초하여 화자 프로파일을 식별하도록 더 구성될 수 있다.

[0363] 컴퓨터에 의해 수신된 가입자 식별자는 제1 익명 식별자일 수 있고, 가입자 식별자는 하나 이상의 화자들에 대한 하나 이상의 화자 식별자들과 연관될 수 있고, 각 화자 식별자는 각 화자 프로파일에 대응하는 제2 익명 식별자일 수 있다.

[0364] 하나 이상의 화자 특성들은 연령 특성 및 성별 특성 중 적어도 하나를 포함할 수 있다.

[0365] 컴퓨터는 인바운드 오디오 신호에 제2 기계 학습 모델을 적용함으로써 인바운드 화자의 연령 특성 및 성별 특성 중 적어도 하나를 결정하도록 더 구성될 수 있다.

[0366] 컴퓨터는 인바운드 화자에 대한 화자 프로파일에 저장된 연령 특성에 기초하여 인바운드 화자의 연령을 결정하도록 더 구성될 수 있다.

[0367] 컴퓨터는 콘텐트 서버로부터 인바운드 화자에 대한 적어도 하나의 화자 특성을 나타내는 화자 정보를 수신하고; 적어도 하나의 화자 특성을 인바운드 화자에 대한 화자 프로파일에 저장하고; 화자 프로파일의 적어도 하나의 화자 특성에 적어도 부분적으로 기초하여 화자 프로파일을 식별하도록 더 구성될 수 있다.

[0368] 컴퓨터는 인바운드 오디오 신호와 함께 콘텐트 서버로부터 인바운드 인증 데이터를 수신하고; 유사도 점수가 유사도 임계값을 충족하고 인바운드 인증 데이터가 화자 프로파일에 저장된 예상된 인증 데이터를 충족하는 것에 기초하여 인바운드 화자를 인증하도록 더 구성될 수 있다.

[0369] 제20항에 따른 시스템에 있어서, 인증 데이터는 최종 사용자 디바이스, 화자 정보 및 생체 인증 정보와 연관된 최종 사용자 디바이스 정보, 메타데이터 중 적어도 하나를 포함한다.

[0370] 컴퓨터는 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하고; 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 하나 이상의 음향 특징들에 기초하여 스푸핑 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하도록 더 구성될 수 있다.

[0371] 컴퓨터는 미디어 콘텐트 데이터베이스에서 하나 이상의 화자 특성들에 대응하는 하나 이상의 콘텐트 특성들을 갖는 하나 이상의 미디어 콘텐트 파일들을 식별하도록 더 구성될 수 있다.

[0372] 컴퓨터는 화자의 하나 이상의 화자 특성들 중 연령 특성이 연령 제한 콘텐트의 하나 이상의 콘텐트 특성들의 해당 연령 제한 특성을 충족하는 것에 기초하여 콘텐트 데이터베이스에서 연령 제한 콘텐트를 활성화하도록 더 구성될 수 있다.

[0373] 일 실시예에서, 컴퓨터 구현 방법은, 컴퓨터에 의해, 화자에 대한 하나 이상의 임베딩들을 포함하는 화자와 연관된 화자 프로파일을 획득하는 단계; 컴퓨터에 의해, 하나 이상의 임베딩들에 대한 하나 이상의 성숙 인자들 및 오수락률에 기초하여 화자에 대한 성문에 대한 성숙 레벨을 결정하는 단계; 및 컴퓨터에 의해, 성숙 레벨 및 하나 이상의 성숙 인자들에 따라 화자 프로파일의 하나 이상의 유사도 임계값들을 업데이트하는 단계를 포함한다.

[0374] 방법은, 화자 프로파일을 획득하는 단계가, 컴퓨터에 의해, 화자가 미디어 콘텐트 시스템의 새로운 사용자라고 결정하는 것에 응답하여 화자 프로파일을 생성하는 단계를 더 포함할 수 있다.

[0375] 방법은 화자 프로파일을 획득하는 단계가, 컴퓨터에 의해, 인바운드 오디오 신호에 기계 학습 아키텍처를 적용함으로써 화자에 대한 인바운드 임베딩을 추출하는 단계; 컴퓨터에 의해, 인바운드 임베딩과 성문 사이의 상대적 거리에 기초하여 인바운드 임베딩에 대한 유사도 점수를 생성하는 단계; 및 컴퓨터에 의해, 인바운드 임베딩에 대한 유사도 점수가 화자 프로파일의 하나 이상의 유사도 임계값들 중 적어도 하나의 유사도 임계값을 충족한다고 결정하는 것에 응답하여 화자 데이터베이스에 저장된 화자 프로파일을 식별하는 단계를 더 포함할 수 있다.

[0376] 적어도 하나의 임계값은 낮은 유사도 임계값을 포함할 수 있고, 하나 이상의 임계값들은 낮은 유사도 임계값 및 높은 유사도 임계값을 포함할 수 있다.

[0377] 방법은, 컴퓨터에 의해, 인바운드 임베딩에 대한 유사도 점수가 하나 이상의 유사도 임계값들 중 높은 유사도 임계값을 충족한다고 결정하는 것에 응답하여 인바운드 임베딩에 따라 화자에 대한 성문을 업데이트하는 단계를 더 포함할 수 있으며, 컴퓨터는 성문을 업데이트한 후 성문에 대한 성숙 레벨을 결정한다.

[0378] 방법은, 컴퓨터에 의해, 성문을 업데이트하는 데 사용되는 인바운드 임베딩에 기초하여 성문에 대한 하나 이상의 성숙 인자들을 업데이트하는 단계를 더 포함할 수 있으며, 컴퓨터는 하나 이상의 성숙 인자들을 업데이트한 후 성문에 대한 성숙 레벨을 결정한다.

[0379] 방법은, 컴퓨터가 성문의 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여, 컴퓨터에 의해, 화자와 연관된 추가 인바운드 임베딩을 요청하는 프롬프트를 생성하는 단계; 및 각 추가 인바운드 임베딩이 하나 이상의 유사도 임계값들을 충족하는 경우, 컴퓨터에 의해, 추가 인바운드 임베딩에 따라 성문을 업데이트하는 단계; 및 컴퓨터에 의해, 각 추가 인바운드 임베딩에 기초하여 성문에 대한 하나 이상의 성숙 인자들을 업데이트하는 단계를 더 포함할 수 있다.

[0380] 방법은 하나 이상의 유사도 임계값들을 업데이트하는 단계가, 컴퓨터에 의해, 성문에 대한 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 화자 프로파일의 하나 이상의 임계값들을 증가시키는 단계를 더 포함할 수 있다.

[0381] 하나 이상의 성숙 인자들은 화자의 임베딩 수, 하나 이상의 임베딩들에서 발생하는 순 음성의 지속시간, 및 하나 이상의 임베딩들로부터의 오디오 품질 중 적어도 하나를 포함할 수 있다.

[0382] 방법은, 컴퓨터에 의해, 관리 컴퓨터로부터 수신된 구성 입력에 기초하여 오수락률을 결정하는 단계를 더 포함할 수 있다.

[0383] 일 실시예에서, 시스템은 복수의 화자 프로파일들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 프로파일 데이터베이스; 및 컴퓨터로서, 화자에 대한 하나 이상의 임베딩들을 포함하는 화자와 연관된 화자 프로파일을 획득하고; 하나 이상의 임베딩들에 대한 하나 이상의 성숙 인자들 및 오수락률에 기초하여 화자에 대한 성문에 대한 성숙 레벨을 결정하고; 성숙 레벨 및 하나 이상의 성숙 인자들에 따라 화자 프로파일의 하나 이상의 유사도 임계값들을 업데이트하도록 구성된 프로세서를 포함하는, 상기 컴퓨터를 포함한다.

[0384] 컴퓨터는, 화자 프로파일을 획득하기 위해, 화자가 미디어 콘텐트 시스템의 새로운 사용자라고 결정하는 것에 응답하여 화자 프로파일을 생성하도록 더 구성될 수 있다.

[0385] 컴퓨터는, 화자 프로파일을 획득하기 위해, 인바운드 오디오 신호에 기계 학습 아키텍처를 적용함으로써 화자에 대한 인바운드 임베딩을 추출하고; 인바운드 임베딩과 성문 사이의 상대적 거리에 기초하여 인바운드 임베딩에 대한 유사도 점수를 생성하고; 인바운드 임베딩에 대한 유사도 점수가 화자 프로파일의 하나 이상의 유사도 임계값들 중 적어도 하나의 유사도 임계값을 충족한다고 결정하는 것에 응답하여 화자 데이터베이스에 저장된 화자 프로파일을 식별하도록 더 구성될 수 있다.

[0386] 적어도 하나의 임계값은 낮은 유사도 임계값을 포함할 수 있고, 하나 이상의 임계값들은 낮은 유사도 임계값 및 높은 유사도 임계값을 포함할 수 있다.

[0387] 컴퓨터는 인바운드 임베딩에 대한 유사도 점수가 하나 이상의 유사도 임계값들 중 높은 유사도 임계값을 충족한다고 결정하는 것에 응답하여 인바운드 임베딩에 따라 화자에 대한 성문을 업데이트하도록 더 구성될 수 있으며, 컴퓨터는 성문을 업데이트한 후 성문에 대한 성숙 레벨을 결정한다.

[0388] 컴퓨터는 성문을 업데이트하는 데 사용되는 인바운드 임베딩에 기초하여 성문에 대한 하나 이상의 성숙 인자들을 업데이트하도록 더 구성될 수 있으며, 컴퓨터는 하나 이상의 성숙 인자들을 업데이트한 후 성문에 대한 성숙 레벨을 결정한다.

[0389] 컴퓨터는, 컴퓨터가 성문의 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여, 화자와 연관된 추가 인바운드 임베딩을 요청하는 프롬프트를 생성하고; 각 추가 인바운드 임베딩이 하나 이상의 유사도 임계값들을 충족하는 경우, 추가 인바운드 임베딩에 따라 성문을 업데이트하고; 각 추가 인바운드 임베딩에 기초하여 성문에 대한 하나 이상의 성숙 인자들을 업데이트하도록 더 구성될 수 있다.

[0390] 컴퓨터는, 하나 이상의 유사도 임계값들 디바이스를 업데이트하기 위해, 성문에 대한 성숙 레벨이 성숙 임계값을 충족한다고 결정하는 것에 응답하여 화자 프로파일의 하나 이상의 임계값들을 증가시키도록 더 구성될 수 있다.

[0391] 하나 이상의 성숙 인자들은 화자의 임베딩 수, 하나 이상의 임베딩들에서 발생하는 순 음성의 지속시간, 및 하나 이상의 임베딩들로부터의 오디오 품질 중 적어도 하나를 포함할 수 있다.

[0392] 컴퓨터는 관리 컴퓨터로부터 수신된 구성 입력에 기초하여 오수락률을 결정하도록 더 구성될 수 있다.

[0393] 일 실시예에서, 디바이스 구현 방법은, 디바이스에 의해, 인바운드 화자의 발화를 포함하는 인바운드 오디오 신호를 수신하는 단계; 디바이스에 의해, 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하는 단계; 디바이스에 의해, 인바운드 임베딩과 비일시적 기계 판독 가능 매체에 저장된 하나 이상의 성문들 사이의 상대적 거리에 따라 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하는 단계; 컴퓨터에 의해, 성문을 사용하여 생성된 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여 인바운드 화자에 대한 성문과 연관된 화자 식별자를 식별하는 단계; 디바이스에 의해, 화자 식별자를 콘텐트 서버에 전송하는 단계를 포함한다.

[0394] 디바이스는 스마트 TV, 셋톱 박스, 에지 디바이스, 리모콘, 및 이동 통신 디바이스 중 적어도 하나를 포함할 수 있다.

[0395] 방법은, 디바이스에 의해, 디바이스에 결합된 디스플레이 디바이스를 통해 디스플레이하기 위해 콘텐트 서버로

부터 미디어 콘텐트를 수신하는 단계를 더 포함할 수 있다.

[0396] 디바이스는 화자 식별자와 함께 미디어 콘텐트에 대한 요청을 콘텐트 서버에 전송할 수 있다.

[0397] 방법은, 디바이스에 의해, 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하는 단계를 더 포함할 수 있으며, 하나 이상의 유형의 생체 인식 특징들을 갖는 특징들은 적어도 하나의 오디오 특징을 포함한다.

[0398] 방법은, 디바이스의 마이크에 의해, 복수의 인바운드 화자들로부터 복수의 발화들을 포함하는 오디오 파형들을 수신하는 단계; 및 디바이스에 의해, 오디오 파형들을 복수의 발화들을 포함하는 인바운드 오디오 신호로 변환하는 단계를 더 포함할 수 있다.

[0399] 인바운드 오디오 신호는 복수의 인바운드 화자들의 복수의 발화들을 포함할 수 있고, 디바이스는 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들에 대한 복수의 유사도 점수들을 생성할 수 있으며, 복수의 유사도 점수는 디바이스의 비일시적 기계 판독 가능 스토리지에 저장된 복수의 성문들에 따라 생성될 수 있다.

[0400] 방법은 디바이스가 복수의 임베딩들 중 제2 인바운드 임베딩에 대한 복수의 유사도 점수들 중 제2 유사도 점수가 적어도 하나의 유사도 임계값을 충족하지 못한다고 결정하는 것에 응답하여, 디바이스에 의해, 비일시적 기계 판독 가능 스토리지에 제2 인바운드 임베딩 및 제2 화자 식별자를 포함하는 제2 인바운드 화자에 대한 제2 화자 프로파일을 생성하는 단계를 더 포함할 수 있다.

[0401] 방법은, 디바이스에 의해, 인바운드 화자에 대한 하나 이상의 등록 오디오 신호들을 요청하는 인바운드 화자에 디스플레이하기 위한 하나 이상의 등록 프롬프트들을 생성하는 단계; 디바이스에 의해, 하나 이상의 등록 오디오 신호들로부터 추출된 하나 이상의 등록 임베딩들에 기초하여 인바운드 화자에 대한 성문을 생성하는 단계; 디바이스가 성문의 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여, 디바이스에 의해, 제2 등록 오디오 신호를 요청하는 제2 등록 프롬프트를 생성하는 단계; 및 디바이스에 의해, 제2 등록 오디오 신호로부터 추출된 제2 등록 임베딩에 기초하여 인바운드 화자에 대한 성문을 업데이트하는 단계를 더 포함할 수 있다.

[0402] 방법은, 디바이스에 의해, 인바운드 오디오 신호로부터 하나 이상의 음향 특징들을 추출하는 단계; 및 디바이스에 의해, 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 하나 이상의 음향 특징들에 기초하여 스푸핑 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하는 단계를 더 포함할 수 있다.

[0403] 일 실시예에서, 시스템은 화자 프로파일들을 포함하는 데이터 레코드들을 저장하도록 구성된 비일시적 기계 판독 가능 저장 매체를 포함하는 화자 데이터베이스; 및 디바이스로서, 인바운드 화자의 발화를 포함하는 인바운드 오디오 신호를 수신하고; 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 인바운드 화자에 대한 인바운드 임베딩을 추출하고; 인바운드 임베딩과 화자 데이터베이스에 저장된 하나 이상의 성문들 사이의 상대적 거리에 따라 인바운드 임베딩에 대한 하나 이상의 유사도 점수들을 생성하고; 컴퓨터에 의해, 성문을 사용하여 생성된 유사도 점수가 유사도 임계값을 충족한다고 결정하는 것에 응답하여 인바운드 화자에 대한 성문과 연관된 화자 식별자를 식별하도록 구성된 프로세서를 포함하는, 상기 디바이스를 포함한다.

[0404] 디바이스는 스마트 TV, 셋톱 박스, 에지 디바이스, 리모콘, 및 이동 통신 디바이스 중 적어도 하나일 수 있다.

[0405] 디바이스는 디바이스에 결합된 디스플레이 디바이스를 통해 디스플레이하기 위해 위한 콘텐트 서버로부터 미디어 콘텐트를 수신하도록 더 구성될 수 있다.

[0406] 디바이스는 화자 식별자와 함께 미디어 콘텐트에 대한 요청을 콘텐트 서버에 전송하도록 더 구성될 수 있다.

[0407] 디바이스는 인바운드 오디오 신호로부터 하나 이상의 특징들을 추출하도록 더 구성될 수 있으며, 하나 이상의 유형의 생체 인식 특징들을 갖는 특징들은 적어도 하나의 오디오 특징을 포함한다.

[0408] 시스템 또는 디바이스는 복수의 인바운드 화자들로부터 복수의 발화들을 포함하는 오디오 파형들을 수신하도록 구성된 마이크를 더 포함할 수 있다. 디바이스는 오디오 파형들을 복수의 발화들을 포함하는 인바운드 오디오 신호로 변환하도록 더 구성될 수 있다.

[0409] 인바운드 오디오 신호는 복수의 인바운드 화자들의 복수의 발화들을 포함할 수 있다. 디바이스는 인바운드 오디오 신호에 임베딩 추출 모델을 적용하여 복수의 인바운드 화자들에 대응하는 복수의 인바운드 임베딩들에 대한 복수의 유사도 점수들을 생성하도록 더 구성될 수 있으며, 디바이스는 화자 데이터베이스에 저장된 복수의 성문들에 따라 복수의 유사도 점수들을 생성한다.

[0410] 디바이스는, 복수의 임베딩들 중 제2 인바운드 임베딩에 대한 복수의 유사도 점수들 중 제2 유사도 점수가 적어도 하나의 유사도 임계값을 충족하지 못한다고 결정하고; 비일시적 기계 판독 가능 스토리지에 제2 인바운드 임베딩 및 제2 화자 식별자를 포함하는 제2 인바운드 화자에 대한 제2 화자 프로파일을 생성하도록 더 구성될 수 있다.

[0411] 디바이스는, 인바운드 화자에 대한 하나 이상의 등록 오디오 신호들을 요청하는 인바운드 화자에 디스플레이하기 위한 하나 이상의 등록 프롬프트들을 생성하고; 하나 이상의 등록 오디오 신호들로부터 추출된 하나 이상의 등록 임베딩들에 기초하여 인바운드 화자에 대한 성문을 생성하고; 디바이스가 성문의 성숙 레벨이 성숙 임계값을 충족하지 못한다고 결정하는 것에 응답하여, 제2 등록 오디오 신호를 요청하는 제2 등록 프롬프트를 생성하고; 제2 등록 오디오 신호로부터 추출된 제2 등록 임베딩에 기초하여 인바운드 화자에 대한 성문을 업데이트하도록 더 구성될 수 있다.

[0412] 디바이스는 인바운드 오디오 신호로부터 하나 이상의 음향 특징들을 추출하고; 인바운드 오디오 신호가 제2 기계 학습 모델을 적용함으로써 하나 이상의 음향 특징들에 기초하여 스푸핑 조건을 포함할 가능성을 나타내는 스푸핑 점수를 계산하도록 더 구성될 수 있다.

[0413] 본원에 개시된 실시예들과 관련하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 회로들, 및 알고리즘 단계들은 전자 하드웨어, 컴퓨터 소프트웨어, 또는 양자의 조합으로 구현될 수 있다. 하드웨어와 소프트웨어의 이러한 호환성을 명확하게 예시하기 위해, 다양한 예시적인 컴포넌트들, 블록들, 모듈들, 회로들 및 단계들이 일반적으로 기능성의 측면에서 위에서 설명되었다. 이러한 기능성이 하드웨어로 구현되는지 또는 소프트웨어로 구현되는지는 전체 시스템에 부과되는 특정 애플리케이션 및 설계 제약 조건에 따라 다르다. 숙련된 기술자들은 각 특정 애플리케이션에 대해 다양한 방식들로 설명된 기능을 구현할 수 있지만, 이러한 구현 결정이 본 발명의 범위로부터의 이탈을 야기하는 것으로 해석되어서는 안 된다.

[0414] 컴퓨터 소프트웨어로 구현된 실시예들은 소프트웨어, 펌웨어, 미들웨어, 마이크로코드, 하드웨어 기술 언어, 또는 이들의 임의의 조합으로 구현될 수 있다. 코드 세그먼트 또는 기계 실행 가능 명령어들은 프로시저, 기능, 서브프로그램, 프로그램, 루틴, 서브루틴, 모듈, 소프트웨어 패키지, 클래스 또는 명령어들, 데이터 구조들 또는 프로그램 명령문들의 임의의 조합을 나타낼 수 있다. 코드 세그먼트는 정보, 데이터, 인수, 파라미터 또는 메모리 콘텐트를 전달 및/또는 수신함으로써 다른 코드 세그먼트 또는 하드웨어 회로에 결합될 수 있다. 정보, 인수, 파라미터, 데이터 등은 메모리 공유, 메시지 전달, 토큰 전달, 네트워크 전송 등을 포함한 임의의 적절한 수단을 통해 전달, 포워딩 또는 전송될 수 있다.

[0415] 이러한 시스템들 및 방법들을 구현하는 데 사용되는 실제 소프트웨어 코드 또는 특수 제어 하드웨어는 본 발명을 제한하지 않는다. 따라서, 시스템들 및 방법들의 동작 및 행동은 소프트웨어 및 제어 하드웨어가 본원의 설명에 기초하여 시스템들 및 방법들을 구현하도록 설계될 수 있다고 이해되는 특정 소프트웨어 코드를 참조하지 않고 설명되었다.

[0416] 소프트웨어로 구현될 때, 기능들은 비일시적 컴퓨터 판독 가능 또는 프로세서 판독 가능 저장 매체에 하나 이상의 명령어들 또는 코드로서 저장될 수 있다. 본원에 개시된 방법 또는 알고리즘의 단계들은 컴퓨터 판독 가능 또는 프로세서 판독 가능 저장 매체에 상주할 수 있는 프로세서 실행 가능 소프트웨어 모듈로 구현될 수 있다. 비일시적 컴퓨터 판독 가능 또는 프로세서 판독 가능 매체는 한 장소에서 다른 장소로의 컴퓨터 프로그램의 전송을 용이하게 하는 컴퓨터 저장 매체 및 유형의 저장 매체를 모두 포함한다. 비일시적 프로세서 판독 가능 저장 매체는 컴퓨터에서 액세스할 수 있는 모든 사용 가능한 매체일 수 있다. 예를 들어, 제한 없이, 이러한 비일시적 프로세서 판독 가능 매체는 RAM, ROM, EEPROM, CD-ROM 또는 기타 광 디스크 스토리지, 자기 디스크 스토리지 또는 기타 자기 스토리지 디바이스들, 또는 명령어들이나 데이터 구조들의 형태로 원하는 프로그램 코드를 저장하는 데 사용될 수 있으며 컴퓨터 또는 프로세서에 의해 액세스될 수 있는 임의의 기타 유형의 저장 매체를 포함할 수 있다. 본원에 사용된 디스크(disk) 및 디스크(disc)에는 디스크(disk)들이 일반적으로 자기적으로 데이터를 재생하는 반면 디스크(disc)들은 레이저들로 데이터를 광학적으로 재생하는 컴팩트 디스크(CD), 레이저 디스크, 광 디스크, 디지털 다목적 디스크(DVD), 플로피 디스크 및 블루-레이 디스크가 포함된다. 위의 조합들도 컴퓨터 판독 가능 매체의 범위에 포함되어야 한다. 추가로, 방법 또는 알고리즘의 동작들은 컴퓨터 프로그램 제품에 통합될 수 있는, 비일시적 프로세서 판독 가능 매체 및/또는 컴퓨터 판독 가능 매체에 하나 또는 임의의 조합 또는 코드 및/또는 명령어 세트로 존재할 수 있다.

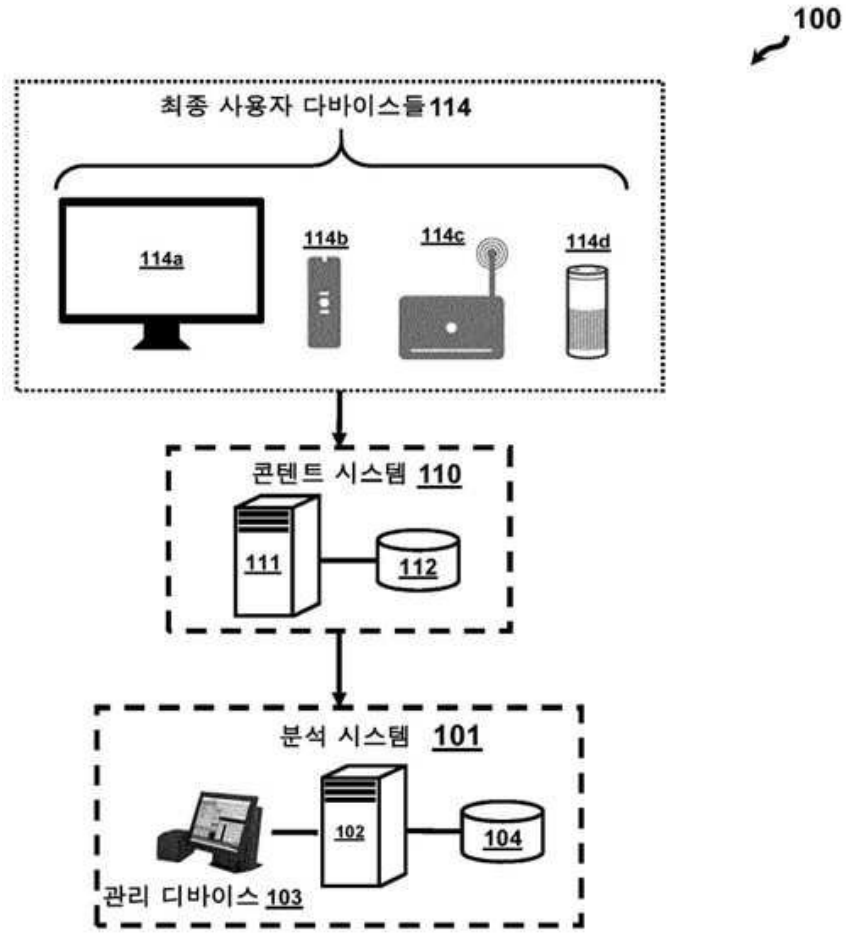
[0417] 개시된 실시예들의 앞선 설명은 당업자가 본 발명을 만들거나 사용할 수 있도록 제공된다. 이들 실시예들에 대한 다양한 수정들은 당업자에게 용이하게 명백할 것이며, 본원에 정의된 일반적인 원리들은 본 발명의 사상 또

는 범위를 벗어나지 않고 다른 실시예들에 적용될 수 있다. 따라서, 본 발명은 본원에 도시된 실시예들로 제한되는 것으로 의도되지 않고, 이하의 청구범위 및 본원에 개시된 원리들 및 신규한 특징들과 일치하는 가장 넓은 범위가 부여되어야 한다.

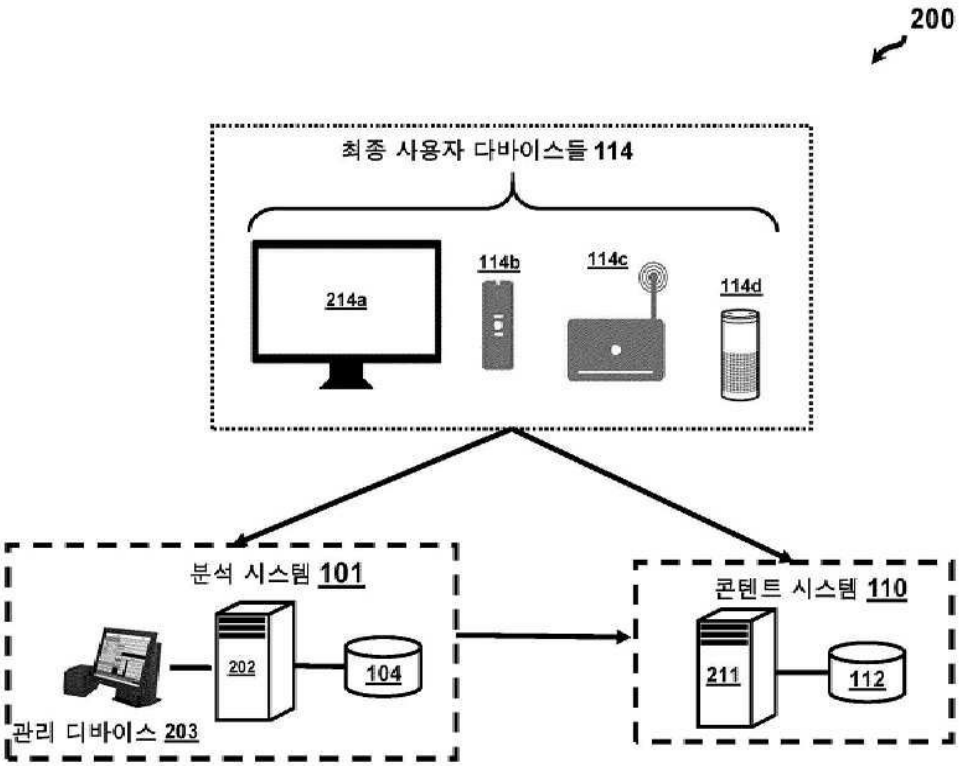
[0418] 다양한 양태들 및 실시예들이 개시되었지만, 다른 양태들 및 실시예들이 고려된다. 개시된 다양한 양태들 및 실시예들은 예시를 위한 것이며 제한하려는 것이 아니며, 진정한 범위 및 사상은 다음 청구범위에 의해 표시된다.

도면

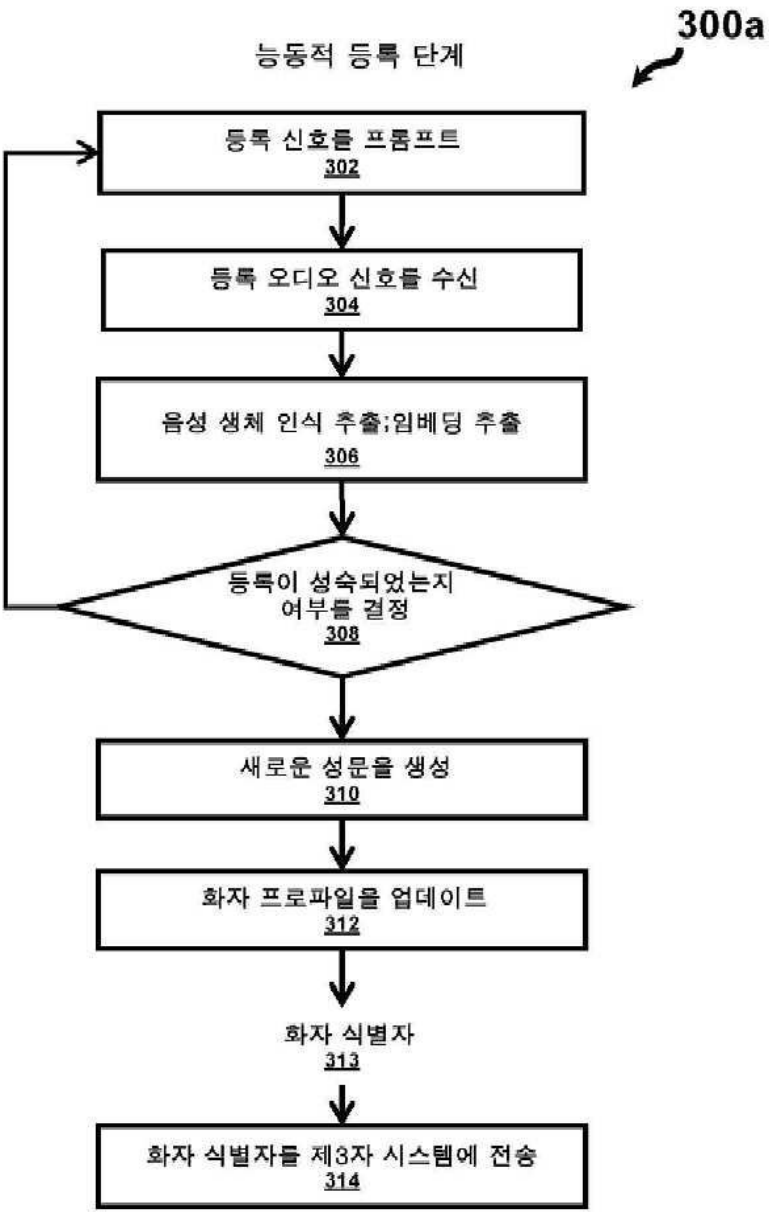
도면1



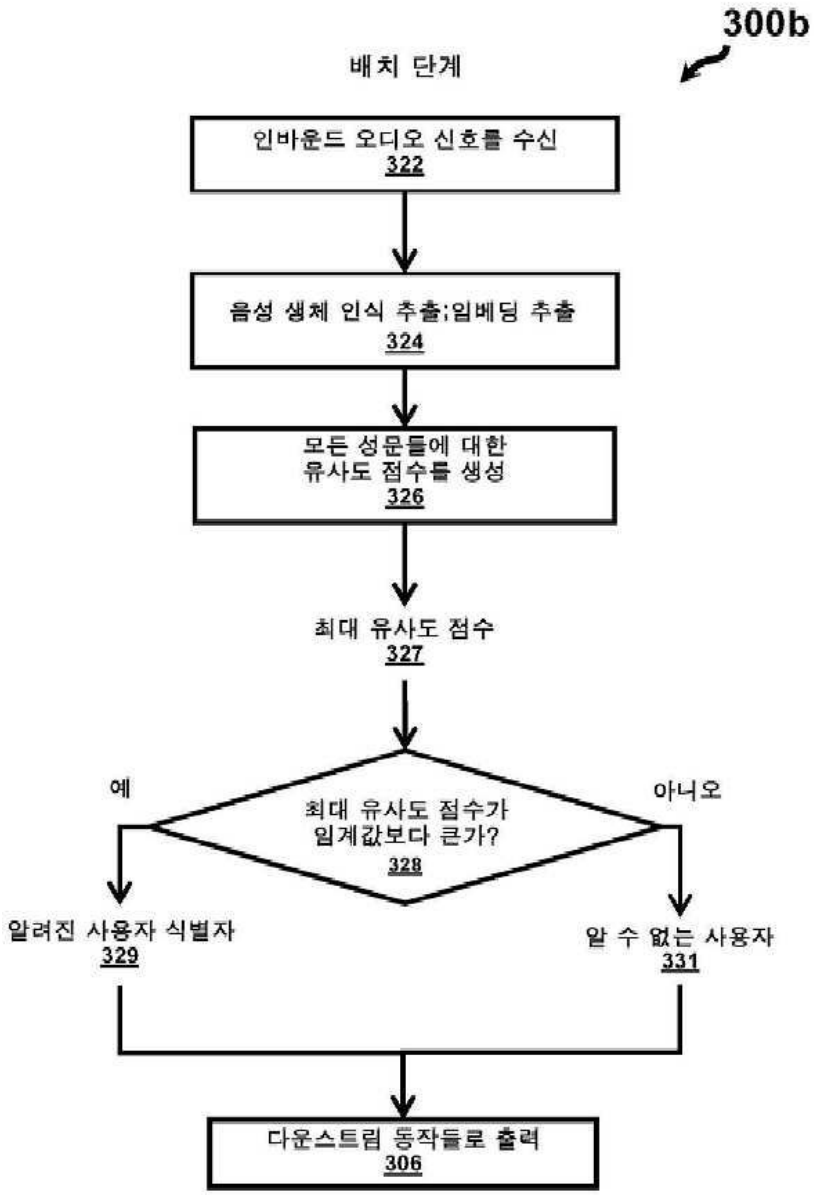
도면2


200
최종 사용자 다바이스들 114
214a
114b
114c
114d
분석 시스템 101
202
104
관리 디바이스 203
콘텐트 시스템 110
211
112

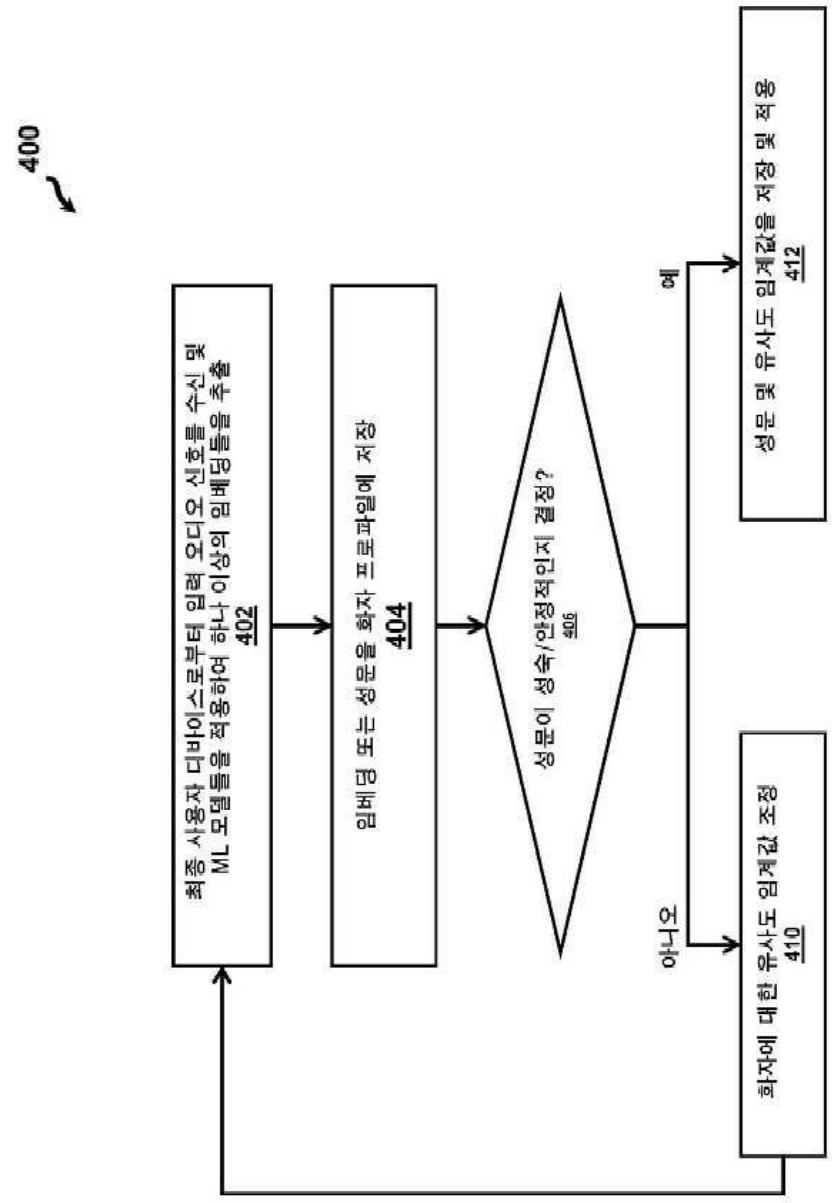
도면3a


300a
능동적 등록 단계
등록 신호를 프롬프트
302
등록 오디오 신호를 수신
304
음성 생체 인식 추출;임베딩 추출
306
등록이 성숙되었는지
여부를 결정
308
새로운 성문을 생성
310
화자 프로파일을 업데이트
312
화자 식별자
313
화자 식별자를 제3자 시스템에 전송
314

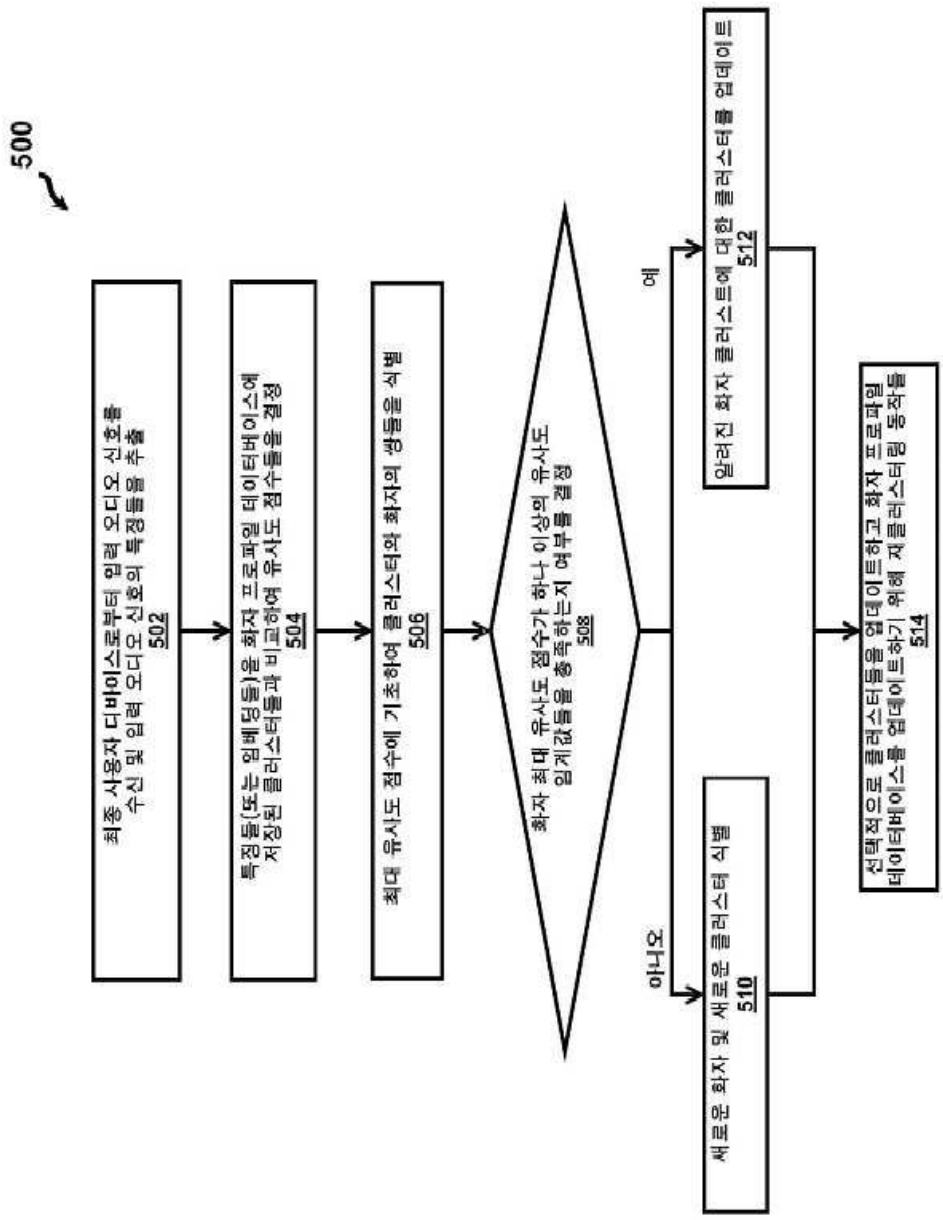
도면3b


300b
배치 단계
인바운드 오디오 신호를 수신
322
음성 생체 인식 추출;임베딩 추출
324
모든 성문들에 대한
유사도 점수를 생성
326
최대 유사도 점수
327
최대 유사도 점수가
임계값보다 큰가?
328
예
아니오
알려진 사용자 식별자
329
알 수 없는 사용자
331
다운스트림 동작들로 출력
306

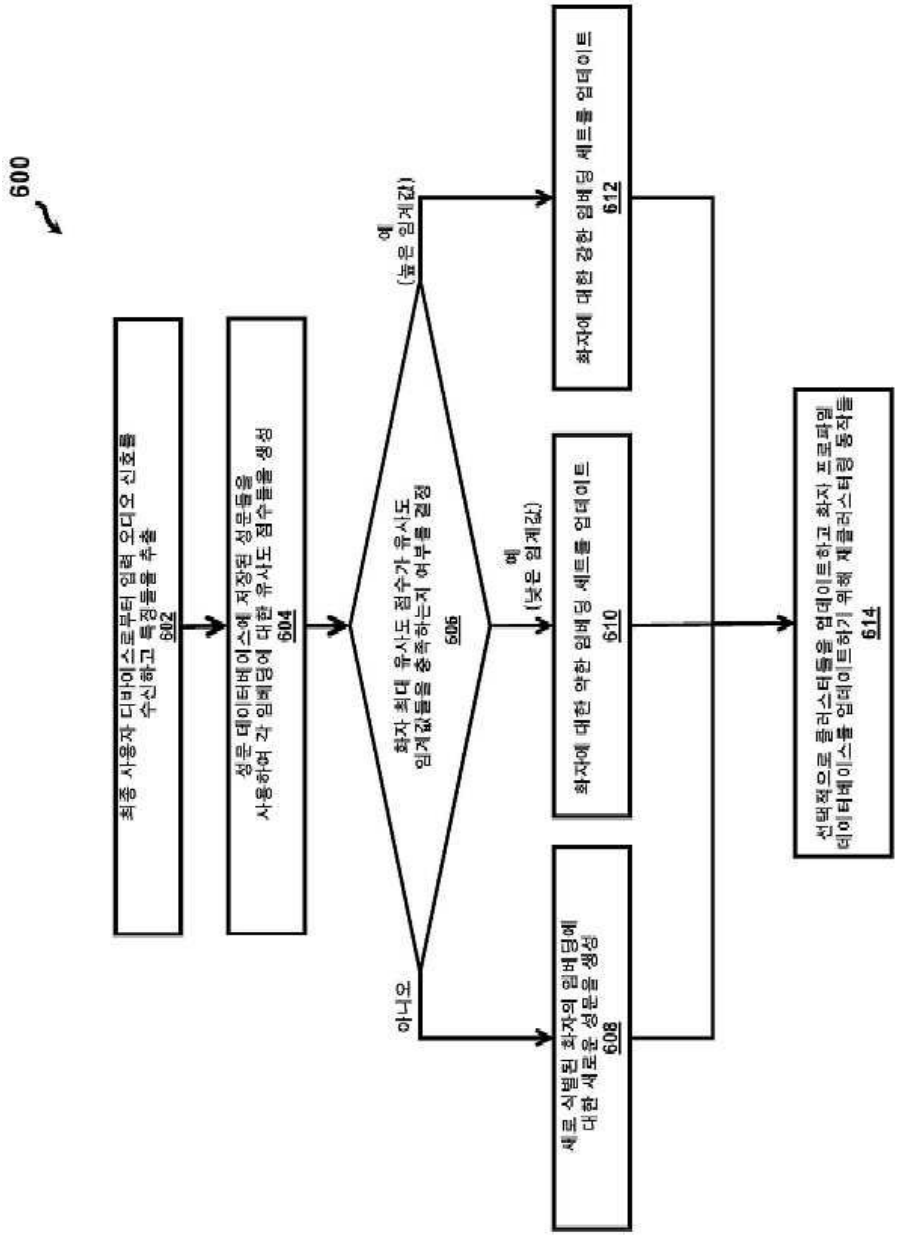
도면4


400
최종 사용자 디바이스로부터 입력 오디오 신호를 수신 및 ML 모델들을 적용하여 하나 이상의 임베딩들을 추출
402
임베딩 또는 성문을 화자 프로파일에 저장
404
성문이 성숙/안정적인지 결정?
406
예
성문 및 유사도 임계값을 저장 및 적용
412
아니오
화자에 대한 유사도 임계값 조정
410

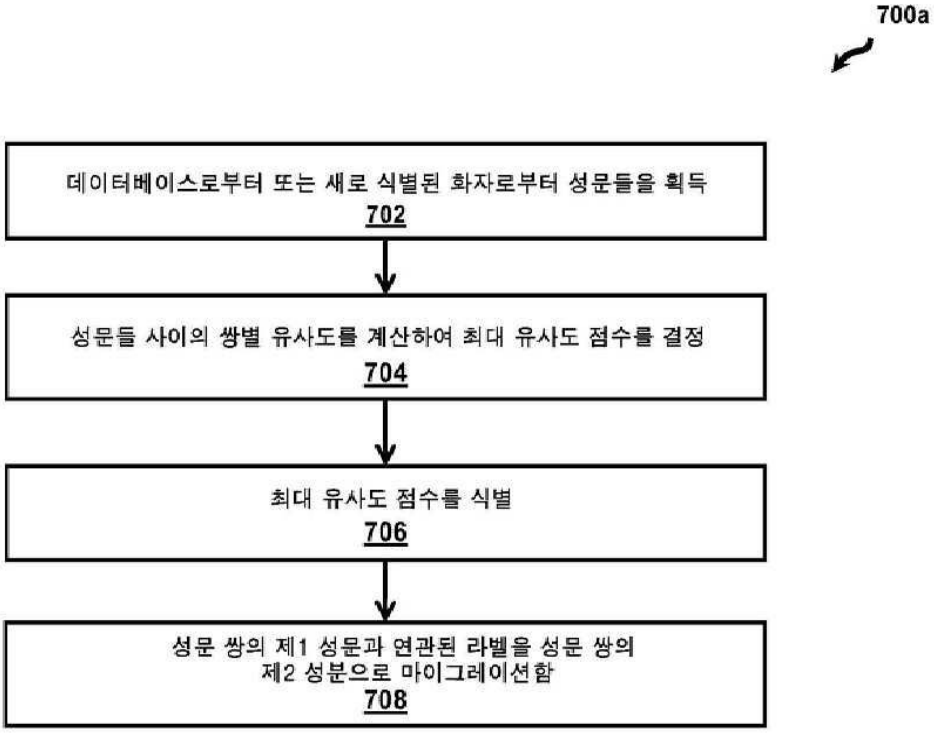
도면5


500
최종 사용자 디바이스로부터 입력 오디오 신호를 수신 및 입력 오디오 신호의 특징들을 추출 502
특징들(또는 임베딩들)을 화자 프로파일 데이터베이스에 저장된 클러스터들과 비교하여 유사도 점수들을 결정 504
최대 유사도 점수에 기초하여 클러스터와 화자의 쌍들을 식별 506
화자 최대 유사도 점수가 하나 이상의 유사도 임계값들을 충족하는지 여부를 결정 508
예
아니오
알려진 화자 클러스트에 대한 클러스터를 업데이트 512
새로운 화자 및 새로운 클러스터 식별 510
선택적으로 클러스터들을 업데이트하고 화자 프로파일 데이터베이스를 업데이트하기 위해 저클러스터링 동작들 514

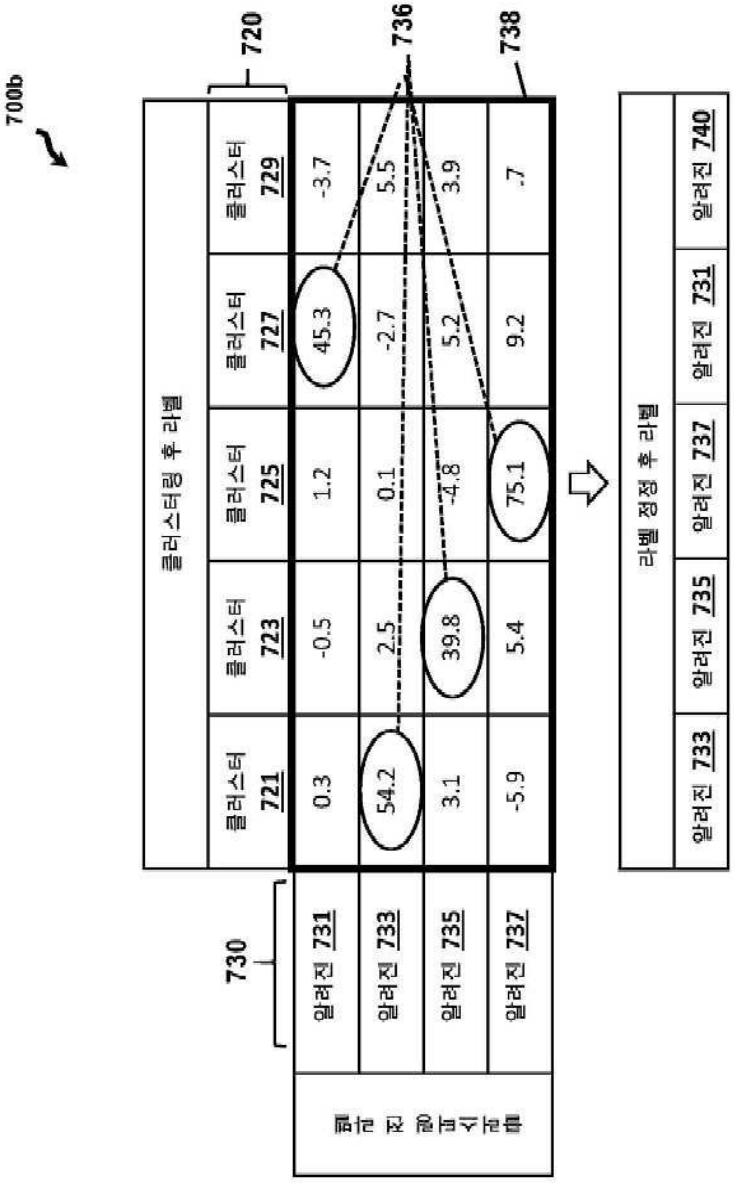
도면6


600
최종 사용자 디바이스로부터 입력 오디오 신호를 수신하고 특징들을 추출 602
성문 데이터베이스에 저장된 성문들을 사용하여 각 임베딩에 대한 유사도 점수들을 생성 604
최자 최대 유사도 점수가 유사도 임계값들을 충족하는지 여부를 결정 606
예 (높은 임계값)
예 (낮은 임계값)
아니오
최자에 대한 강한 임베딩 세트를 업데이트 612
최자에 대한 약한 임베딩 세트를 업데이트 610
새로 식별된 최자의 임베딩에 대한 새로운 성문을 생성 608
선택적으로 클러스터들을 업데이트하고 최자 프로파일 데이터베이스를 업데이트하기 위해 재클러스터링 동작들 614

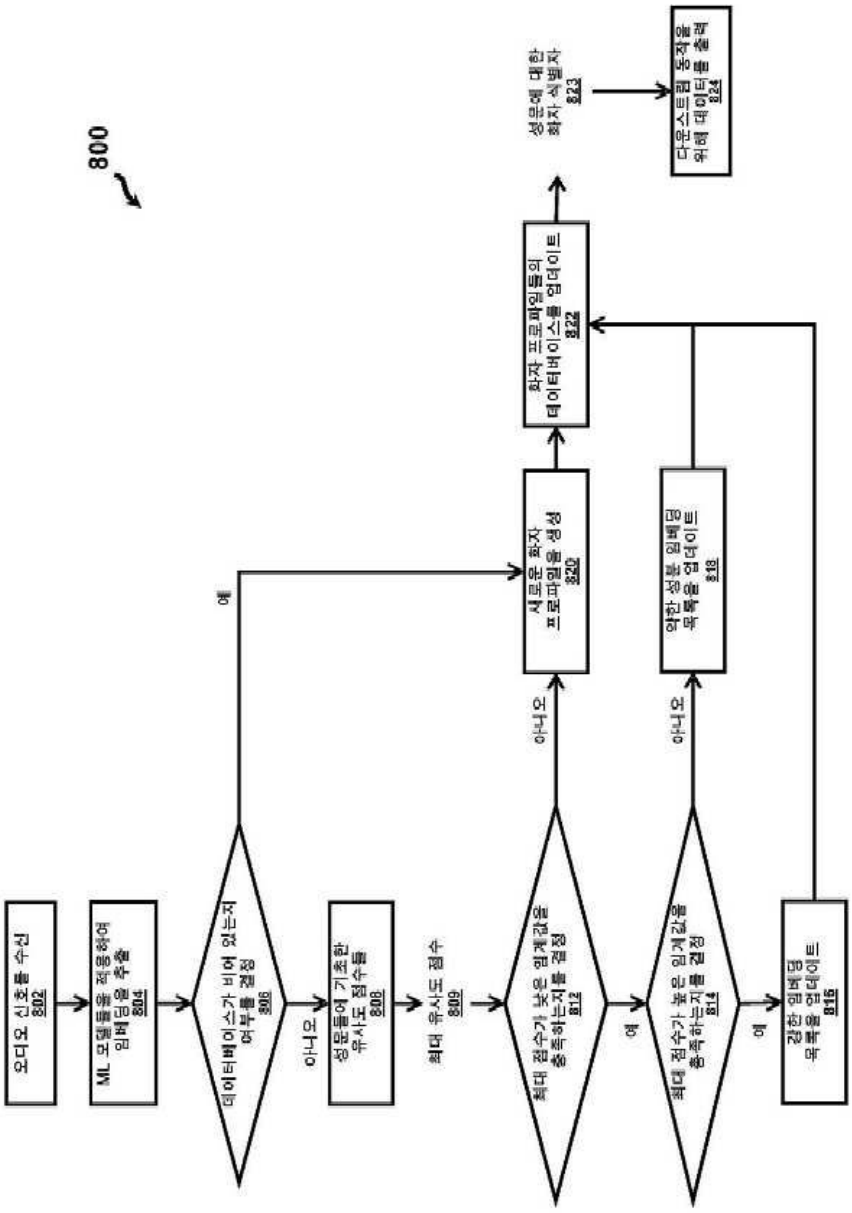
도면7a


700a
데이터베이스로부터 또는 새로 식별된 화자로부터 성문들을 획득
702
성문들 사이의 쌍별 유사도를 계산하여 최대 유사도 점수를 결정
704
최대 유사도 점수를 식별
706
성문 쌍의 제1 성문과 연관된 라벨을 성문 쌍의 제2 성분으로 마이그레이션함
708

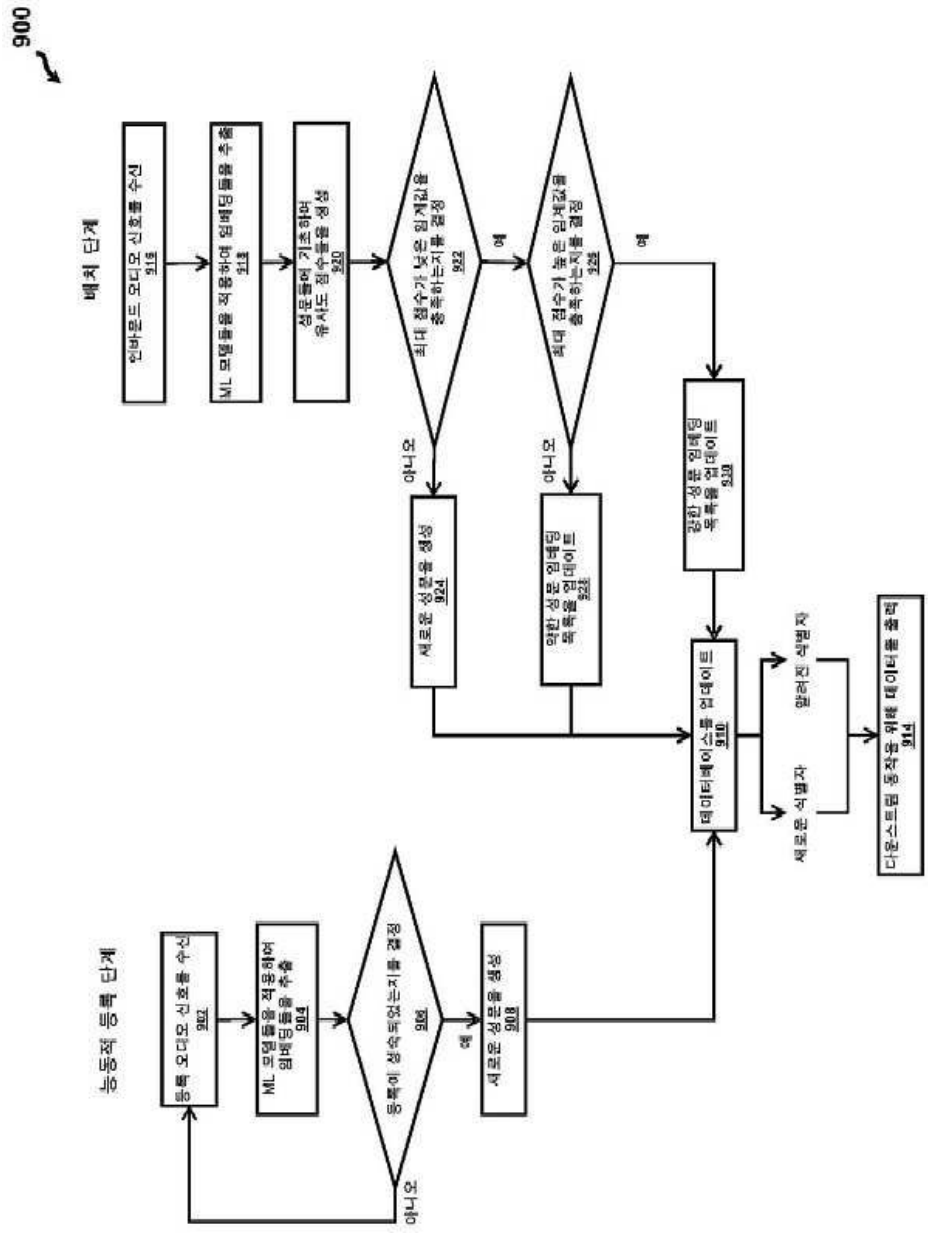
도면7b


700b
720
736
738
730
클러스터링 후 라벨
클러스터 721
클러스터 723
클러스터 725
클러스터 727
클러스터 729
알려진 731
알려진 733
알려진 735
알려진 737
0.3
-0.5
1.2
45.3
-3.7
54.2
2.5
0.1
-2.7
5.5
3.1
39.8
-4.8
5.2
3.9
-5.9
5.4
75.1
9.2
.7
라벨 정정 후 라벨
알려진 733
알려진 735
알려진 737
알려진 731
알려진 740

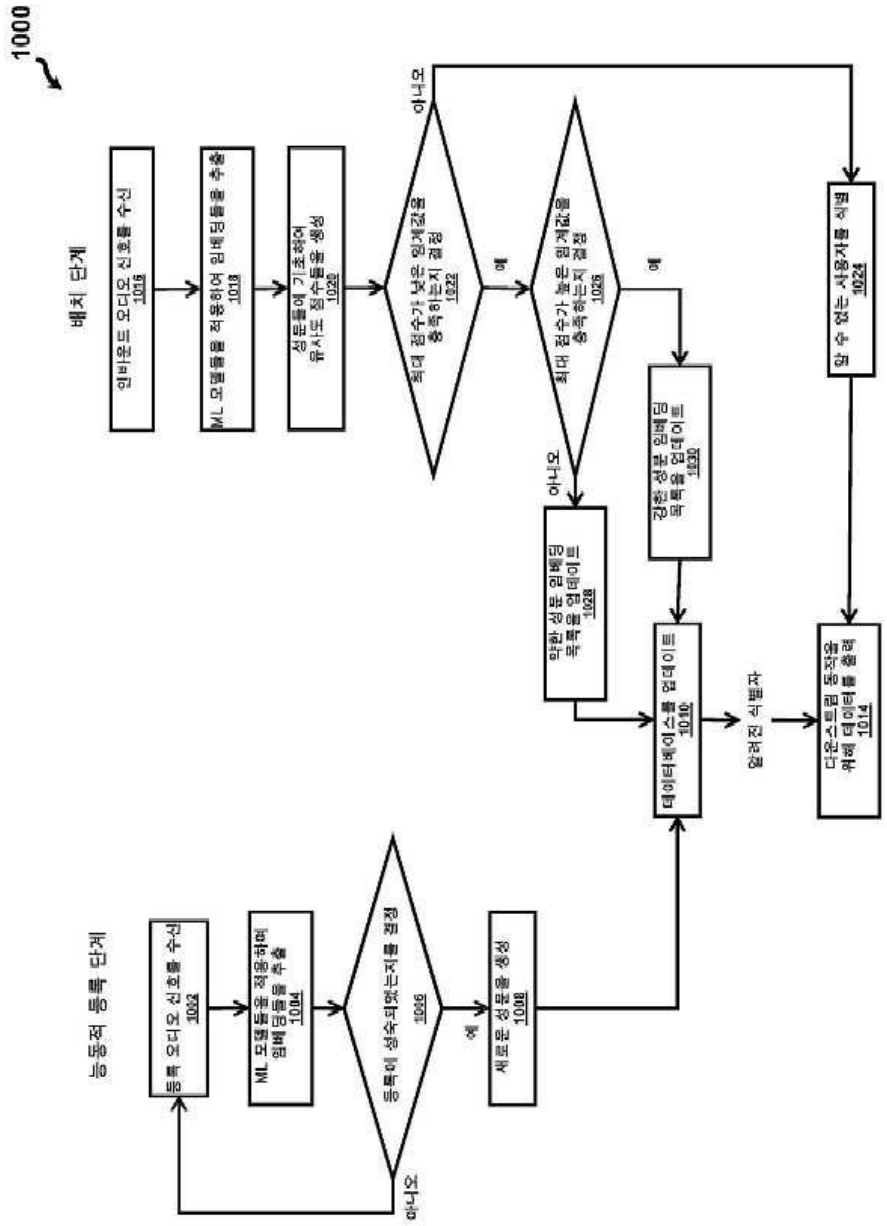
도면8


800
오디오 신호를 수신
802
ML 모델들을 적용하여
임베딩을 추출
804
데이터베이스가 비어 있는지
여부를 결정
806
예
아니오
질문들에 기초한
유사도 점수들
808
최대 유사도 점수
809
최대 점수가 낮은 임계값을
충족하는지를 결정
812
아니오
새로운 화자
프로파일을 생성
820
예
최대 점수가 높은 임계값을
충족하는지를 결정
814
아니오
알려진 화자 임베딩
목록을 업데이트
818
예
강한 임베딩
목록을 업데이트
816
화자 프로파일들의
데이터베이스를 업데이트
822
성문에 대한
화자 식별자
823
다운스트림 동작을
위해 데이터를 출력
824

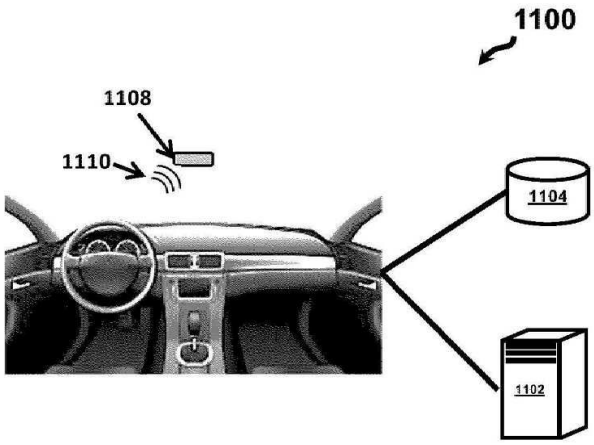
도면9


900
배치 단계
인바운드 오디오 신호를 수신
912
ML 모델들을 적용하여 임베딩들을 추출
913
성문들에 기초하여
유사도 점수들을 생성
920
최대 점수가 낮은 임계값을
충족하는지를 결정
922
예
최대 점수가 높은 임계값을
충족하는지를 결정
926
예
아니오
아니오
새로운 성문을 생성
924
약한 성문 임베딩
목록을 업데이트
928
강한 성문 임베딩
목록을 업데이트
930
데이터베이스를 업데이트
910
알려진 식별자
새로운 식별자
다운스트림 동작을 위해 데이터를 출력
914
능동적 등록 단계
등록 오디오 신호를 수신
902
ML 모델들을 적용하여
임베딩들을 추출
904
등록이 성숙되었는지를 결정
906
예
새로운 성문을 생성
908
아니오

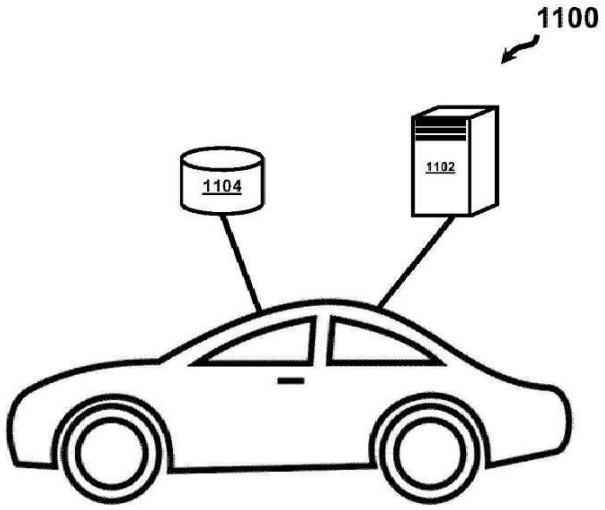
도면10


1000
배치 단계
인바운드 오디오 신호를 수신
1016
ML 모델들을 적용하여 임베딩들을 추출
1018
성문들에 기초하여
유사도 점수들을 생성
1020
최대 점수가 낮은 임계값을
충족하는지 결정
1022
아니오
예
최대 점수가 높은 임계값을
충족하는지 결정
1026
아니오
예
알 수 없는 사용자를 식별
1024
강화 성문 임베딩
목록을 업데이트
1030
약한 성문 임베딩
목록을 업데이트
1028
데이터베이스를 업데이트
1010
업데이트된 식별자
다운스트림 동작을
위해 데이터를 출력
1014
능동적 등록 단계
등록 오디오 신호를 수신
1002
ML 모델들을 적용하여
임베딩들을 추출
1004
등록이 성숙되었는지를 결정
1006
아니오
예
새로운 성문을 생성
1008

도면11a


1100
1108
1110
1104
1102

도면11b



도면12

1200

1202 1204 1206

등록 수	임계값@ FAR = 0.5%	임계값@ FAR = 5%
1	3.36	-0.5
2	3.83	-0.2
3	4.02	-0.08
4	4.13	0.02
5	4.20	0.05
6	4.21	0.06
7	4.29	0.10
8	4.34	0.12
9	4.35	0.13
10	4.37	0.15

1210
1214
1208
1212