(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0027475
(43) 공개일자 2020년03월12일

(51) 국제특허분류(Int. Cl.)
G10L 21/013 (2013.01) *G10L 15/02* (2006.01)
G10L 15/06 (2006.01) *G10L 15/22* (2006.01)
G10L 19/018 (2013.01) *G10L 25/30* (2013.01)
(52) CPC특허분류
G10L 21/013 (2013.01)
G10L 15/02 (2013.01)
(21) 출원번호 10-2019-7038068
(22) 출원일자(국제) 2018년05월24일
심사청구일자 2020년03월05일
(85) 번역문제출일자 2019년12월23일
(86) 국제출원번호 PCT/US2018/034485
(87) 국제공개번호 WO 2018/218081
국제공개일자 2018년11월29일
(30) 우선권주장
62/510,443 2017년05월24일 미국(US)

(71) 출원인
모듈레이트, 인크
미국, 매사추세츠 02142, 캠브리지, 14층, 원 브로드웨이
(72) 발명자
허프먼, 윌리엄 카터
미국, 매사추세츠 02141, 캠브리지, 에이피티. 106, 세컨드 스트리트 110
파파스, 마이클
미국, 매사추세츠 02141, 캠브리지, 에이피티. 106, 세컨드 스트리트 110
(74) 대리인
특허법인(유한) 다래

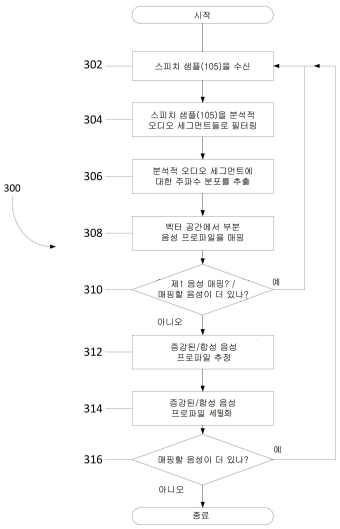
전체 청구항 수 : 총 78 항

(54) 발명의 명칭 **음성 대 음성 변환을 위한 시스템 및 방법**

(57) 요 약

음성 변환 시스템을 구축하는 방법은 타겟 음성으로부터의 타겟 정보 및 소스 스피치 데이터를 사용한다. 방법은 소스 스피치 데이터 및 음색 공간 내에 있는 타겟 음색 데이터를 수신한다. 생성기는 소스 스피치 데이터 및 타겟 음색 데이터의 함수로서 제1 후보 데이터를 생성한다. 판별기는 복수의 상이한 음성의 음색 데이터를 참조하여 제1 후보 데이터를 타겟 음색 데이터와 비교한다. 판별기는 제1 후보 데이터와 타겟 음색 데이터 사이의 불일치들을 결정한다. 판별기는 불일치들에 관한 정보를 포함하는 불일치 메시지를 생성한다. 불일치 메시지는 생성기에 피드백되고, 생성기는 제2 후보 데이터를 생성한다. 음색 공간에서의 타겟 음색 데이터는 피드백의 결과로서 생성기 및/또는 판별기에 의해 생성된 정보를 사용하여 세밀화된다.

대 표 도 - 도3



(52) CPC특허분류
G10L 15/063 (2013.01)
G10L 15/22 (2013.01)
G10L 19/018 (2013.01)
G10L 25/30 (2013.01)
G10L 2015/025 (2013.01)
G10L 2021/0135 (2013.01)

명 세 서

청구범위

청구항 1

타겟 음성으로부터의 타겟 음성 정보, 및 소스 음성의 스피치 세그먼트를 나타내는 스피치 데이터를 사용하여 스피치 변환 시스템을 구축하는 방법으로서,

소스 음성의 제1 스피치 세그먼트를 나타내는 소스 스피치 데이터를 수신하는 단계;

상기 타겟 음성에 관한 타겟 음색 데이터를 수신하는 단계로서, 상기 타겟 음색 데이터는 음색 공간 내에 있는, 단계;

생성적 기계 학습 시스템을 사용하여, 상기 소스 스피치 데이터 및 상기 타겟 음색 데이터의 함수로서 제1 후보 음성 내의 제1 후보 스피치 세그먼트를 나타내는 제1 후보 스피치 데이터를 생성하는 단계;

판별적 기계 학습 시스템을 사용하여, 복수의 상이한 음성의 음색 데이터를 참조하여 상기 제1 후보 음색 데이터를 상기 타겟 음색 데이터와 비교하는 단계로서,

상기 판별적 기계 학습 시스템을 사용하는 것은 상기 복수의 상이한 음성의 상기 음색 데이터를 참조하여 상기 제1 후보 스피치 데이터와 상기 타겟 음색 데이터 사이의 적어도 하나의 불일치를 결정하는 것을 포함하며, 상기 판별적 기계 학습 시스템은 상기 제1 후보 스피치 데이터와 상기 타겟 음색 데이터 사이의 상기 불일치에 관한 정보를 갖는 불일치 메시지를 생성하는, 단계;

상기 불일치 메시지를 상기 생성적 기계 학습 시스템에 피드백하는 단계;

상기 생성적 기계 학습 시스템을 사용하여, 상기 불일치 메시지의 함수로서 제2 후보 음성 내의 제2 후보 스피치 세그먼트를 나타내는 제2 후보 스피치 데이터를 생성하는 단계; 및

상기 피드백의 결과로서 상기 생성적 기계 학습 시스템 및/또는 판별적 기계 학습 시스템에 의해 생성된 정보를 사용하여 상기 음색 공간에서 상기 타겟 음색 데이터를 세밀화하는 단계

를 포함하는, 방법.

청구항 2

제1항에 있어서, 상기 소스 스피치 데이터는 상기 소스 음성의 오디오 입력으로부터 유래되는, 방법.

청구항 3

제1항에 있어서, 상기 제2 후보 스피치 세그먼트는 상기 제1 후보 스피치 세그먼트보다 높은 상기 타겟 음성으로부터 유래될 확률을 제공하는, 방법.

청구항 4

제1항에 있어서, 상기 소스 스피치 데이터를 상기 타겟 음색으로 변환하는 단계를 더 포함하는, 방법.

청구항 5

제1항에 있어서, 상기 타겟 음색 데이터는 상기 타겟 음성 내의 오디오 입력으로부터 획득되는, 방법.

청구항 6

제1항에 있어서, 상기 기계 학습 시스템은 신경망인, 방법.

청구항 7

제1항에 있어서,

벡터 공간에서 상기 복수의 음성 및 상기 제1 후보 음성의 표현을 각각의 음성에 의해 제공되는 상기 스피치 세

그먼트 내의 주파수 분포의 함수로서 매핑하는 단계를 더 포함하는, 방법.

청구항 8

제7항에 있어서,

상기 불일치 메시지의 함수로서 상기 제2 후보 음성을 반영하기 위해 상기 벡터 공간에서 상기 복수의 음성의 표현들에 대해 상기 제1 후보 음성의 표현을 조정하는 단계를 더 포함하는, 방법.

청구항 9

제1항에 있어서, 상기 불일치 메시지는 상기 판별적 신경망이 상기 제1 후보 음성이 상기 타겟 음성이라는 95 퍼센트 미만의 신뢰 구간을 가질 때 생성되는, 방법.

청구항 10

제1항에 있어서,

상기 후보 음성을 상기 복수의 음성과 비교함으로써 상기 후보 음성에 아이덴티티를 할당하는 단계를 더 포함하는, 방법.

청구항 11

제1항에 있어서, 상기 복수의 음성은 벡터 공간 내에 있는, 방법.

청구항 12

제1항에 있어서, 상기 타겟 음색 데이터는 시간 수용 필드에 의해 필터링되는, 방법.

청구항 13

제1항에 있어서, 상기 생성적 기계 학습 시스템을 사용하여, 널 불일치 메시지의 함수로서 최종 후보 음성에서 최종 후보 스피치 세그먼트를 생성하는 단계를 더 포함하며,

상기 최종 후보 스피치 세그먼트는 상기 타겟 음색 내의 상기 제1 스피치 세그먼트를 모방하는, 방법.

청구항 14

제13항에 있어서, 상기 시간 수용 필드는 약 10 밀리초 내지 약 1000 밀리초사이인, 방법.

청구항 15

제1항에 있어서, 타겟 스피치 세그먼트로부터 상기 타겟 음색 데이터를 추출하기 위한 수단을 더 포함하는, 방법.

청구항 16

스피치 변환 시스템을 트레이닝하기 위한 시스템으로서,

소스 음성의 제1 스피치 세그먼트를 나타내는 소스 스피치 데이터;

타겟 음성에 관한 타겟 음색 데이터;

상기 소스 스피치 데이터 및 상기 타겟 음색 데이터의 함수로서 제1 후보 음성 내의 제1 후보 스피치 세그먼트를 나타내는 제1 후보 스피치 데이터를 생성하도록 구성된 생성적 기계 학습 시스템;

판별적 기계 학습 시스템

을 포함하고, 상기 판별적 기계 학습 시스템은:

복수의 상이한 음성의 음색 데이터를 참조하여 상기 제1 후보 스피치 데이터를 상기 타겟 음색 데이터와 비교하고,

상기 복수의 상이한 음성의 상기 음색 데이터를 참조하여 상기 제1 후보 스피치 데이터와 상기 타겟 음색 데이

터 사이에 적어도 하나의 불일치가 존재하는지를 결정하고, 상기 적어도 하나의 불일치가 존재할 때:

상기 제1 후보 스피치 데이터와 상기 타겟 음색 데이터 사이의 상기 불일치에 관한 정보를 갖는 불일치 메시지를 생성하며,

상기 불일치 메시지를 다시 상기 생성적 기계 학습 시스템에 제공하도록 구성되는, 시스템.

청구항 17

제16항에 있어서, 상기 생성적 기계 학습 시스템은 상기 불일치 메시지의 함수로서 제2 후보 스피치 세그먼트를 생성하도록 구성되는, 시스템.

청구항 18

제16항에 있어서, 상기 기계 학습 시스템은 신경망인, 시스템.

청구항 19

제16항에 있어서,

상기 후보 음성을 포함하는 상기 복수의 음성의 표현을 각각의 음성에 의해 제공되는 상기 스피치 세그먼트 내의 주파수 분포의 함수로서 매핑하도록 구성되는 벡터 공간을 더 포함하는, 시스템.

청구항 20

제19항에 있어서, 음성 특징 추출기가 상기 벡터 공간에서 상기 복수의 음성의 표현들에 대해 상기 후보 음성의 표현을 조정하여, 상기 불일치 메시지의 함수로서 제2 후보 음성을 업데이트 및 반영하도록 구성되는, 시스템.

청구항 21

제16항에 있어서, 상기 후보 음성은 상기 판별적 신경망이 95 퍼센트 미만의 신뢰 구간을 가질 때 상기 타겟 음성으로부터 구별되는, 시스템.

청구항 22

제16항에 있어서, 상기 판별적 기계 학습 시스템은 상기 제1 또는 제2 후보 음성을 상기 복수의 음성과 비교함으로써 상기 후보 음성의 화자의 아이덴티티를 결정하도록 구성되는, 시스템.

청구항 23

제16항에 있어서, 복수의 음성을 포함하도록 구성되는 벡터 공간을 더 포함하는, 시스템.

청구항 24

제16항에 있어서, 상기 생성적 기계 학습 시스템은 널 불일치 메시지의 함수로서 최종 후보 음성에서 최종 후보 스피치 세그먼트를 생성하도록 구성되며,

상기 최종 후보 스피치 세그먼트는 상기 제1 스피치 세그먼트를 상기 타겟 음성으로서 모방하는, 시스템.

청구항 25

제16항에 있어서, 상기 타겟 음색 데이터는 시간 수용 필드에 의해 필터링되는, 시스템.

청구항 26

제25항에 있어서, 상기 시간 수용 필드는 약 10 밀리초 내지 약 2000 밀리초사이인, 시스템.

청구항 27

제16항에 있어서, 상기 소스 스피치 데이터는 소스 오디오 입력으로부터 유래되는, 시스템.

청구항 28

타겟 음성 음색을 갖는 출력 음성으로 변환하기 위한 소스 음성으로부터의 스피치 세그먼트를 나타내는 소스 스

피치 데이터를 사용하여 스피치 변환 시스템을 트레이닝하기 위해 컴퓨터 시스템 상에서 사용하기 위한 컴퓨터 프로그램 제품으로서, 상기 컴퓨터 프로그램 제품은 컴퓨터 판독가능 프로그램 코드를 갖는 유형적인 비일시적 컴퓨터 사용가능 매체를 포함하고, 상기 컴퓨터 판독가능 프로그램 코드는:

생성적 기계 학습 시스템으로 하여금 상기 소스 스피치 데이터 및 타겟 음색 데이터의 함수로서 제1 후보 음성 내의 제1 후보 스피치 세그먼트를 나타내는 제1 후보 스피치 데이터를 생성하게 하는 프로그램 코드;

판별적 기계 학습 시스템으로 하여금 복수의 상이한 음성의 음색 데이터를 참조하여 상기 제1 후보 스피치 데이터를 상기 타겟 음색 데이터와 비교하게 하는 프로그램 코드;

상기 판별적 기계 학습 시스템으로 하여금 상기 복수의 상이한 음성의 상기 음색 데이터를 참조하여 상기 제1 후보 스피치 데이터와 상기 타겟 음색 데이터 사이의 적어도 하나의 불일치를 결정하게 하는 프로그램 코드;

상기 판별적 기계 학습 시스템으로 하여금 상기 복수의 상이한 음성의 상기 음색 데이터를 참조하여 상기 제1 후보 스피치 데이터와 상기 타겟 음색 데이터 사이의 상기 불일치에 관한 정보를 갖는 불일치 메시지를 생성하게 하는 프로그램 코드;

상기 판별적 기계 학습 시스템으로 하여금 상기 불일치 메시지를 다시 상기 생성적 기계 학습 시스템에 제공하게 하는 프로그램 코드; 및

상기 생성적 기계 학습 시스템으로 하여금 상기 불일치 메시지의 함수로서 제2 후보 음성 내의 제2 후보 스피치 세그먼트를 나타내는 제2 후보 스피치 데이터를 생성하게 하는 프로그램 코드

를 포함하는, 컴퓨터 프로그램 제품.

청구항 29

제28항에 있어서,

타겟 오디오 입력으로부터 상기 타겟 음색 데이터를 추출하는 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 30

제28항에 있어서, 상기 기계 학습 시스템은 신경망인, 컴퓨터 프로그램 제품.

청구항 31

제28항에 있어서,

벡터 공간에서 상기 복수의 음성 및 상기 후보 음성 각각의 표현을 각각의 음성으로부터의 상기 음색 데이터의 함수로서 매핑하기 위한 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 32

제31항에 있어서,

상기 벡터 공간에서 상기 복수의 음성의 적어도 하나의 표현에 대해 상기 후보 음성의 상기 표현을 조정하여, 상기 불일치 메시지의 함수로서 상기 제2 후보 음성을 업데이트 및 반영하기 위한 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 33

제28항에 있어서,

상기 후보 음성을 상기 복수의 음성과 비교함으로써 상기 후보 음성에 화자 아이덴티티를 할당하기 위한 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 34

제28항에 있어서,

시간 수용 필드를 사용하여, 입력된 타겟 오디오를 필터링하여 상기 음색 데이터를 생성하기 위한 프로그램 코

드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 35

제34항에 있어서, 상기 시간 수용 필드는 약 10 밀리초 내지 약 2000 밀리초사이인, 컴퓨터 프로그램 제품.

청구항 36

제28항에 있어서,

상기 소스 음성으로부터의 상기 스피치 세그먼트를 나타내는 상기 소스 스피치 데이터를 상기 타겟 음색 내의 변환된 스피치 세그먼트로 변환하기 위한 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 37

제36항에 있어서,

상기 변환된 스피치 세그먼트에 워터마크를 추가하기 위한 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 38

음색 벡터 공간을 구축하기 위한 음색 벡터 공간 구축 시스템으로서,

a) 제1 음성 내의 제1 음색 데이터를 포함하는 제1 스피치 세그먼트 및 b) 제2 음성 내의 제2 음색 데이터를 포함하는 제2 스피치 세그먼트를 수신하도록 구성된 입력;

상기 제1 스피치 세그먼트를 제1 복수의 더 작은 분석적 오디오 세그먼트로 변환하는 시간 수용 필드로서, 상기 제1 복수의 더 작은 분석적 오디오 세그먼트 각각은 상기 제1 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 갖고, 필터는 또한 상기 시간 수용 필드를 사용하여 상기 제2 스피치 세그먼트를 제2 복수의 더 작은 분석적 오디오 세그먼트로 변환하도록 구성되고, 상기 제2 복수의 더 작은 분석적 오디오 세그먼트 각각은 상기 제2 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 가지는, 시간 수용 필드;

a) 상기 제1 스피치 세그먼트로부터의 상기 제1 복수의 분석적 오디오 세그먼트 및 b) 상기 제2 스피치 세그먼트로부터의 상기 제2 복수의 분석적 오디오 세그먼트의 주파수 분포의 함수로서 상기 음색 벡터 공간에서 상기 제2 음성에 대해 상기 제1 음성을 매핑하도록 구성된 기계 학습 시스템

을 포함하는, 시스템.

청구항 39

제38항에 있어서, 데이터베이스는 제3 음성 내의 제3 스피치 세그먼트를 수신하도록 구성되고,

상기 기계 학습 시스템은 시간 수용 필드를 사용하여 상기 제3 스피치 세그먼트를 복수의 더 작은 분석적 오디오 세그먼트로 필터링하고, 상기 벡터 공간에서 상기 제1 음성 및 상기 제2 음성에 대해 상기 제3 음성을 매핑하도록 구성되는, 시스템.

청구항 40

제39항에 있어서, 상기 제1 음성 및 상기 제2 음성에 대해 상기 제3 음성을 매핑하는 것은 상기 벡터 공간 내의 상기 제2 음성에 대한 상기 제1 음성의 상대적 위치를 변경하는, 시스템.

청구항 41

제38항에 있어서, 상기 시스템은 적어도 하나의 음성에서 영어의 각각의 인간 음소를 매핑하도록 구성되는, 시스템.

청구항 42

제38항에 있어서, 상기 수용 필드는 상기 음성의 스피치 레이트 및/또는 액센트를 캡처하지 못하도록 충분히 작은, 시스템.

청구항 43

제38항에 있어서, 상기 시간 수용 필드는 약 10 밀리초 내지 약 2000 밀리초사이인, 시스템.

청구항 44

스피치 세그먼트들을 변환하기 위한 음색 벡터 공간을 구축하는 방법으로서,

a) 제1 음성 내의 음색 데이터를 포함하는 제1 스피치 세그먼트 및 b) 제2 음성 내의 음색 데이터를 포함하는 제2 스피치 세그먼트를 수신하는 단계;

시간 수용 필드를 사용하여, 상기 제1 스피치 세그먼트 및 상기 제2 스피치 세그먼트 각각을 복수의 더 작은 분석적 오디오 세그먼트로 필터링하는 단계로서, 각각의 분석적 오디오 세그먼트는 상기 음색 데이터를 나타내는 주파수 분포를 가지는, 단계;

기계 학습 시스템을 사용하여, 상기 제1 스피치 세그먼트 및 상기 제2 스피치 세그먼트로부터의 상기 복수의 분석적 오디오 세그먼트 중 적어도 하나에서의 상기 주파수 분포의 함수로서 벡터 공간에서 상기 제2 음성에 대해 상기 제1 음성을 매핑하는 단계

를 포함하는, 방법.

청구항 45

제44항에 있어서, 상기 제1 스피치 세그먼트 및 상기 제2 스피치 세그먼트 각각을 필터링하기 위한 수단을 더 포함하는, 방법.

청구항 46

제44항에 있어서, 상기 제2 음성에 대해 상기 제1 음성을 매핑하기 위한 수단을 더 포함하는, 방법.

청구항 47

제44항에 있어서, 상기 필터링은 기계 학습 시스템에 의해 수행되는, 방법.

청구항 48

제44항에 있어서,

제3 음성 내의 제3 스피치 세그먼트를 수신하는 단계;

시간 수용 필드를 사용하여, 상기 제3 스피치 세그먼트를 복수의 더 작은 분석적 오디오 세그먼트로 필터링하는 단계; 및

상기 벡터 공간에서 상기 제1 음성 및 상기 제2 음성에 대해 상기 제3 음성을 매핑하는 단계

를 더 포함하는, 방법.

청구항 49

제48항에 있어서,

상기 제3 음성의 매핑의 함수로서 상기 벡터 공간 내의 상기 제2 음성에 대한 상기 제1 음성의 상대적 위치를 조정하는 단계를 더 포함하는, 방법.

청구항 50

제48항에 있어서, 상기 수용 필드는 상기 음성의 스피치 레이트 및/또는 액센트를 캡처하지 못하도록 충분히 작은, 방법.

청구항 51

제48항에 있어서,

적어도 하나의 음성에서 영어의 각각의 인간 음소를 매핑하는 단계를 더 포함하는, 방법.

청구항 52

제48항에 있어서, 상기 시간 수용 필드는 약 10 밀리초 내지 약 500 밀리초사이인, 방법.

청구항 53

음성들을 저장하고 조직화하기 위해 컴퓨터 시스템 상에서 사용하기 위한 컴퓨터 프로그램 제품으로서, 상기 컴퓨터 프로그램 제품은 컴퓨터 판독가능 프로그램 코드를 갖는 유형적인 비일시적 컴퓨터 사용가능 매체를 포함하고, 상기 컴퓨터 판독가능 프로그램 코드는:

입력으로 하여금, a) 제1 음성 내의 음색 데이터를 포함하는 제1 스피치 세그먼트 및 b) 제2 음성 내의 음색 데이터를 포함하는 제2 음성을 수신하게 하는 프로그램 코드;

시간 수용 필드를 사용하여, 상기 제1 스피치 세그먼트 및 상기 제2 스피치 세그먼트 각각을 복수의 더 작은 분석적 오디오 세그먼트로 필터링하는 프로그램 코드로서, 각각의 분석적 오디오 세그먼트는 상기 음색 데이터를 나타내는 주파수 분포를 가지는, 프로그램 코드; 및

기계 학습 시스템으로 하여금, 상기 제1 스피치 세그먼트 및 상기 제2 스피치 세그먼트로부터의 상기 복수의 분석적 오디오 세그먼트 중 적어도 하나에서의 상기 주파수 분포의 함수로서 벡터 공간에서 상기 제2 음성에 대해 상기 제1 음성을 매핑하게 하는 프로그램 코드

를 포함하는, 컴퓨터 프로그램 제품.

청구항 54

제53항에 있어서, 상기 제1 스피치 세그먼트 및 상기 제2 스피치 세그먼트 각각을 필터링하기 위한 수단을 더 포함하는, 컴퓨터 프로그램 제품.

청구항 55

제53항에 있어서, 상기 벡터 공간에서 상기 제2 음성에 대해 상기 제1 음성을 매핑하기 위한 수단을 더 포함하는, 컴퓨터 프로그램 제품.

청구항 56

제53항에 있어서,

입력으로 하여금, c) 제3 음성 내의 제3 스피치 세그먼트를 수신하게 하는 프로그램 코드; 및

시간 수용 필드를 사용하여, 상기 제3 스피치 세그먼트를 복수의 더 작은 분석적 오디오 세그먼트로 필터링하기 위한 프로그램 코드

를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 57

제56항에 있어서,

상기 벡터 공간에서 상기 제1 음성 및 상기 제2 음성에 대해 상기 제3 음성을 매핑하기 위한 프로그램 코드를 더 포함하며,

상기 제1 음성 및 상기 제2 음성에 대해 상기 제3 음성을 매핑하는 것은 상기 벡터 공간 내의 상기 제2 음성에 대한 상기 제1 음성의 상대적 위치를 변경하는, 컴퓨터 프로그램 제품.

청구항 58

제53항에 있어서,

상기 음성의 스피치 레이트 및/또는 액센트를 캡처하지 못하도록 상기 시간 수용 필드를 정의하도록 구성되는 프로그램 코드를 더 포함하는, 컴퓨터 프로그램 제품.

청구항 59

제58항에 있어서, 상기 시간 수용 필드는 약 10 밀리초 내지 약 500 밀리초사이인, 컴퓨터 프로그램 제품.

청구항 60

음색 벡터 공간을 구축하기 위한 음색 벡터 공간 구축 시스템으로서,

a) 제1 음성 내의 제1 음색 데이터를 포함하는 제1 스피치 세그먼트 및 b) 제2 음성 내의 제2 음색 데이터를 포함하는 제2 스피치 세그먼트를 수신하도록 구성된 입력;

a) 상기 제1 스피치 세그먼트를 상기 제1 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 갖는 제1 복수의 더 작은 분석적 오디오 세그먼트로 필터링하고, b) 상기 제2 스피치 세그먼트를 제2 복수의 더 작은 분석적 오디오 세그먼트로 필터링하기 위한 수단으로서, 상기 제2 복수의 더 작은 분석적 오디오 세그먼트 각각은 상기 제 2 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 가지는, 수단;

a) 상기 제1 스피치 세그먼트로부터의 상기 제1 복수의 분석적 오디오 세그먼트 및 b) 상기 제2 스피치 세그먼트로부터의 상기 제2 복수의 분석적 오디오 세그먼트의 상기 주파수 분포의 함수로서 상기 음색 벡터 공간에서 상기 제2 음성에 대해 상기 제1 음성을 매핑하기 위한 수단

을 포함하는, 음색 벡터 공간 구축 시스템.

청구항 61

음색 벡터 공간을 사용하여 새로운 음색을 갖는 새로운 음성을 구축하는 방법으로서,

시간 수용 필드를 사용하여 필터링된 음색 데이터를 수신하는 단계로서, 상기 음색 데이터는 상기 음색 벡터 공간에서 매핑되고, 타겟 음색 데이터는 복수의 상이한 음성과 관련되고, 상기 복수의 상이한 음성 각각은 상기 음색 벡터 공간에서 각각의 음색 데이터를 가지는, 단계; 및

기계 학습 시스템을 사용하여, 상기 복수의 상이한 음성의 상기 타겟 음색 데이터를 사용하여 상기 새로운 음색을 구축하는 단계

를 포함하는, 방법.

청구항 62

제61항에 있어서, 상기 타겟 음색 데이터를 필터링하기 위한 수단을 더 포함하는, 방법.

청구항 63

제61항에 있어서,

소스 스피치를 제공하는 단계; 및

소스 케이던스 및 소스 액센트를 유지하면서 상기 소스 스피치를 상기 새로운 음색으로 변환하는 단계

를 더 포함하는, 방법.

청구항 64

제61항에 있어서,

새로운 음성으로부터 새로운 스피치 세그먼트를 수신하는 단계;

신경망을 사용하여 상기 새로운 스피치 세그먼트를 새로운 분석적 오디오 세그먼트로 필터링하는 단계;

복수의 매핑된 음성에 대해 상기 벡터 공간에서 상기 새로운 음성을 매핑하는 단계; 및

상기 복수의 매핑된 음성에 대한 상기 새로운 음성의 관계에 기초하여 상기 새로운 음성의 특성들 중 적어도 하나를 결정하는 단계

를 더 포함하는, 방법.

청구항 65

제61항에 있어서,

생성적 신경망을 사용하여, 제1 음성과 제2 음성 사이의 수학적 연산의 함수로서, 후보 음성에서 제1 후보 스피치 세그먼트를 생성하는 단계를 더 포함하는, 방법.

청구항 66

제61항에 있어서, 상기 벡터 공간 내의 음성 표현들의 클러스터는 특정 액센트를 나타내는, 방법.

청구항 67

제61항에 있어서, 상기 복수의 음성 각각으로부터의 스피치 세그먼트는 상이한 스피치 세그먼트인, 방법.

청구항 68

음색 벡터 공간을 사용하여 새로운 타겟 음성을 생성하는 시스템으로서,

시간 수용 필드를 사용하여 통합된 음색 데이터를 저장하도록 구성된 음색 벡터 공간;

시간 수용 필드를 사용하여 필터링된 음색 데이터로서, 상기 음색 데이터는 복수의 상이한 음성과 관련된, 음색 데이터; 및

상기 음색 데이터를 사용하여 상기 음색 데이터를 상기 새로운 타겟 음성으로 변환하도록 구성된 기계 학습 시스템

을 포함하는, 시스템.

청구항 69

제68항에 있어서, 상기 기계 학습 시스템은 신경망인, 시스템.

청구항 70

제68항에 있어서, 상기 기계 학습 시스템은:

새로운 음성으로부터 새로운 스피치 세그먼트를 수신하고,

상기 새로운 스피치 세그먼트를 새로운 음색 데이터로 필터링하고,

복수의 음색 데이터에 대해 상기 벡터 공간에서 상기 새로운 음색 데이터를 매핑하고,

상기 복수의 음색 데이터에 대한 상기 새로운 음색 데이터의 관계에 기초하여 상기 새로운 음성의 적어도 하나의 음성 특성을 결정하도록 구성되는, 시스템.

청구항 71

제68항에 있어서, 상기 음색 데이터를 상기 새로운 타겟 음성으로 변환하는 것은 상기 음색 데이터의 적어도 하나의 음성 특성을 변수로서 사용하여 수학적 연산을 수행함으로써 개시되는, 시스템.

청구항 72

제68항에 있어서, 상기 벡터 공간 내의 음성 표현들의 클러스터는 특정 액센트를 나타내는, 시스템.

청구항 73

음색 벡터 공간을 사용하여 새로운 타겟 음성을 생성하기 위해 컴퓨터 시스템에서 사용하기 위한 컴퓨터 프로그램 제품으로서, 상기 컴퓨터 프로그램 제품은 컴퓨터 판독가능 프로그램 코드를 갖는 유형적인 비일시적 컴퓨터 사용가능 매체를 포함하고, 상기 컴퓨터 판독가능 프로그램 코드는:

시간 수용 필드를 사용하여 필터링된 음색 데이터를 수신하기 위한 프로그램 코드로서, 상기 음색 데이터는 상기 시간 수용 필드를 통합하는 상기 음색 벡터 공간에 저장되고, 상기 음색 데이터는 복수의 상이한 음성과 관

련된, 프로그램 코드; 및

기계 학습 시스템을 사용하여, 상기 음색 데이터를 상기 음색 데이터를 사용하여 상기 새로운 타겟 음성으로 변환하는 프로그램 코드

를 포함하는, 컴퓨터 프로그램 제품.

청구항 74

제73항에 있어서,

새로운 음성으로부터 새로운 스피치 세그먼트를 수신하기 위한 프로그램 코드;

상기 기계 학습 시스템으로 하여금 상기 새로운 스피치 세그먼트를 새로운 분석적 오디오 세그먼트로 필터링하게 하는 프로그램 코드;

복수의 매핑된 음성에 대해 상기 벡터 공간에서 상기 새로운 음성을 매핑하는 프로그램 코드; 및

상기 복수의 매핑된 음성에 대한 상기 새로운 음성의 관계에 기초하여 상기 새로운 음성의 특성들 중 적어도 하나를 결정하기 위한 프로그램 코드

를 더 포함하는, 프로그램 코드.

청구항 75

제73항에 있어서, 상기 기계 학습 시스템은 신경망인, 프로그램 코드.

청구항 76

제73항에 있어서, 상기 음색 데이터를 상기 새로운 타겟 음성으로 변환하는 것은 상기 음색 데이터의 적어도 하나의 음성 특성을 변수로서 사용하여 수학 연산을 수행함으로써 개시되는, 프로그램 코드.

청구항 77

제73항에 있어서, 상기 벡터 공간 내의 음성 표현들의 클러스터는 특정 액센트를 나타내는, 프로그램 코드.

청구항 78

소스 음색으로부터 타겟 음색으로 스피치 세그먼트를 변환하는 방법으로서,

복수의 상이한 음성에 관련된 음색 데이터를 저장하는 단계로서, 상기 복수의 상이한 음성 각각은 음색 벡터 공간에서 각각의 음색 데이터를 갖고, 상기 음색 데이터는 시간 수용 필드를 사용하여 필터링되고 상기 음색 벡터 공간에서 매핑되는, 단계;

소스 음성으로 변환하기 위해 소스 음성 내에 소스 스피치 세그먼트를 수신하는 단계;

타겟 음성의 선택을 수신하는 단계로서, 상기 타겟 음성은 타겟 음색을 갖고, 상기 타겟 음성은 상기 복수의 상이한 음성을 참조하여 상기 음색 벡터 공간에서 매핑되는, 단계;

기계 학습 시스템을 사용하여, 상기 소스 음성의 음색으로부터의 상기 소스 스피치 세그먼트를 상기 타겟 음성의 음색으로 변환하는 단계

를 포함하는, 방법.

발명의 설명

기 술 분 야

[0001] 우선권

[0002] 본 특허 출원은 2017년 5월 24일자로 출원되었고, 발명의 명칭이 "적대적 신경망을 이용하는 음색 전달 시스템 및 방법(TIMBRE TRANSFER SYSTEMS AND METHODS UTILIZING ADVERSARIAL NEURAL NETWORKS)"이며 윌리엄 시. 허프만(William C. Huffman)을 발명자로 하는 미국 특허 가출원 제62/510,443호로부터 우선권을 주장하며, 그 개시

내용 전체가 본 명세서에 참조로 통합된다.

[0003] 발명의 분야

[0004] 본 발명은 일반적으로 음성 변환에 관한 것으로서, 특히 본 발명은 합성 음성 프로파일들을 생성하는 것에 관한 것이다.

배 경 기 술

[0005] 최근에, 아마존의 알렉사(Alexa), 애플의 시리(Siri) 및 구글의 어시스턴트(Assistant)와 같은 개인용 음성 활성화 어시스턴트들(personal voice-activated assistants)의 사용으로 인해 음성 기술에 대한 관심이 최고조에 이르렀다. 또한, 팟캐스트들 및 오디오북 서비스들도 최근에 대중화되었다.

발명의 내용

[0006] 본 발명의 일 실시예에 따르면, 음성 변환 시스템을 구축하는 방법은 타겟 음성으로부터의 타겟 음성 정보, 및 소스 음성의 스피치 세그먼트(speech segment)를 나타내는 스피치 데이터를 사용한다. 방법은 소스 음성의 제1 스피치 세그먼트를 나타내는 소스 스피치 데이터를 수신한다. 방법은 또한 타겟 음성에 관한 타겟 음색(target timbre) 데이터를 수신한다. 타겟 음색 데이터는 음색 공간 내에 있다. 생성적 기계 학습 시스템(generative machine learning system)이 소스 스피치 데이터 및 타겟 음색 데이터의 함수로서 제1 후보 음성 내의 제1 후보 스피치 세그먼트를 나타내는 제1 후보 스피치 데이터를 생성한다. 판별적 기계 학습 시스템(discriminative machine learning system)이 복수의 상이한 음성의 음색 데이터를 참조하여 제1 후보 스피치 데이터를 타겟 음색 데이터와 비교하는 데 사용된다. 판별적 기계 학습 시스템은 복수의 상이한 음성의 음색 데이터를 참조하여 제1 후보 스피치 데이터와 타겟 음색 데이터 사이의 적어도 하나의 불일치를 결정한다. 판별적 기계 학습 시스템은 또한 제1 후보 스피치 데이터와 타겟 음색 데이터 사이의 불일치에 관한 정보를 갖는 불일치 메시지를 생성한다. 방법은 또한 불일치 메시지를 생성적 기계 학습 시스템에 피드백한다. 생성적 기계 학습 시스템은 불일치 메시지의 함수로서 제2 후보 음성 내의 제2 후보 스피치 세그먼트를 나타내는 제2 후보 스피치 데이터를 생성한다. 음색 공간에서의 타겟 음색 데이터는 피드백의 결과로서 생성적 기계 학습 시스템 및/또는 판별적 기계 학습 시스템에 의해 생성된 정보를 사용하여 세밀화(refine)된다.

[0007] 일부 실시예들에서, 소스 스피치 데이터는 타겟 음색으로 변환된다. 많은 방식들 중에서 특히, 소스 스피치 데이터는 소스 음성으로부터의 오디오 입력으로부터 온 것일 수 있다. 유사한 방식으로, 타겟 음색 데이터는 타겟 음성으로부터의 오디오 입력으로부터 획득될 수 있다. 타겟 음색 데이터는 타겟 스피치 세그먼트로부터 추출될 수 있다. 또한, 타겟 음색 데이터는 시간 수용 필드(temporal receptive field)에 의해 필터링될 수 있다.

[0008] 기계 학습 시스템은 신경망일 수 있고, 복수의 음성은 벡터 공간에 있을 수 있다. 따라서, 복수의 음성 및 제1 후보 음성은 각각의 음성에 의해 제공되는 스피치 세그먼트에서의 주파수 분포의 함수로서 벡터 공간에서 매핑될 수 있다. 또한, 벡터 공간 내의 복수의 음성의 표현들에 대한 제1 후보 음성의 표현은 제2 후보 음성을 불일치 메시지의 함수로서 반영하도록 조정될 수 있다. 따라서, 시스템은 후보 음성을 복수의 음성과 비교함으로써 후보 음성에 아이덴티티를 할당할 수 있다.

[0009] 일부 실시예들에서, 판별적 신경망(discriminative neural network)은 제1 후보 음성이 타겟 음성이라는 95 퍼센트 미만의 신뢰 구간이 존재할 때 불일치 메시지를 생성한다. 따라서, 제2 후보 스피치 세그먼트는 제1 후보 스피치 세그먼트보다 판별기에 의해 타겟 음성으로서 식별될 더 높은 확률을 제공한다. 따라서, 일부 실시예들은 최종 후보 음성에서 최종 후보 스피치 세그먼트를 널 불일치 메시지(null inconsistency message)의 함수로서 생성하기 위해 생성적 기계 학습 시스템을 사용한다. 최종 후보 스피치 세그먼트는 소스 스피치 세그먼트를 모방하지만, 타겟 음색을 갖는다.

[0010] 일 실시예에 따르면, 스피치 변환 시스템을 트레이닝하기 위한 시스템은 소스 음성의 제1 스피치 세그먼트를 나타내는 소스 스피치 데이터를 포함한다. 시스템은 또한 타겟 음성과 관련된 타겟 음색 데이터를 포함한다. 또한, 시스템은 소스 스피치 데이터 및 타겟 음색 데이터의 함수로서 제1 후보 음성 내의 제1 후보 스피치 세그먼트를 나타내는 제1 후보 스피치 데이터를 생성하도록 구성된 생성적 기계 학습 시스템을 포함한다. 시스템은 또한 복수의 상이한 음성의 음색 데이터를 참조하여 제1 후보 스피치 데이터를 타겟 음색 데이터와 비교하도록 구성된 판별적 기계 학습 시스템을 갖는다. 또한, 판별적 기계 학습은 복수의 상이한 음성의 음색 데이터를 참조하여 제1 후보 스피치 데이터와 타겟 음색 데이터 사이에 적어도 하나의 불일치가 존재하는지를 결정하도록

구성된다. 적어도 하나의 불일치가 존재할 때, 판별적 기계 학습은 제1 후보 스피치 데이터와 타겟 음색 데이터 사이의 불일치에 관한 정보를 갖는 불일치 메시지를 생성한다. 또한, 판별적 기계 학습은 불일치 메시지를 생성적 기계 학습 시스템에 다시 제공한다.

[0011] 타겟 음색 데이터는 시간 수용 필드에 의해 필터링된다. 일부 실시예들에서, 시간 수용 필드는 약 10 밀리초(milliseconds) 내지 2000 밀리초 사이이다. 보다 구체적으로, 시간 수용 필드는 약 10 밀리초 내지 1000 밀리초 사이일 수 있다.

[0012] 일부 실시예들에서, 음성 특징 추출기(voice feature extractor)는 벡터 공간 내의 복수의 음성의 표현들에 대해 후보 음성의 표현을 조정하여 불일치 메시지의 함수로서 제2 후보 음성을 업데이트하고 반영하도록 구성된다. 또한, 판별적 기계 학습 시스템은 제1 또는 제2 후보 음성을 복수의 음성과 비교함으로써 후보 음성의 화자의 아이덴티티를 결정하도록 구성될 수 있다.

[0013] 타겟 음색 데이터는 타겟 오디오 입력으로부터 추출될 수 있다. 소스 스피치 데이터는 타겟 음색 내의 변환된 스피치 세그먼트로 변환될 수 있다. 시스템은 또한 변환된 스피치 세그먼트에 워터마크를 추가할 수 있다.

[0014] 본 발명의 또 다른 실시예에 따르면, 음색 벡터 공간을 구축하기 위한 음색 벡터 공간 구축 시스템은 입력을 포함한다. 입력은 a) 제1 음성 내의 제1 음색 데이터를 포함하는 제1 스피치 세그먼트 및 b) 제2 음성 내의 제2 음색 데이터를 포함하는 제2 스피치 세그먼트를 수신하도록 구성된다. 시스템은 또한 제1 스피치 세그먼트를 제1 복수의 더 작은 분석적 오디오 세그먼트(smaller analytical audio segments)로 변환하기 위한 시간 수용 필드를 포함한다. 제1 복수의 더 작은 분석적 오디오 세그먼트 각각은 제1 음색 데이터의 상이한 부분을 나타내는 주파수 분포(frequency distribution)를 갖는다. 필터는 또한 제2 스피치 세그먼트를 제2 복수의 더 작은 분석적 오디오 세그먼트로 변환하기 위해 시간 수용 필드를 사용하도록 구성된다. 제2 복수의 더 작은 분석적 오디오 세그먼트 각각은 제2 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 갖는다. 시스템은 또한 음색 벡터 공간에서 제2 음성에 대해 제1 음성을 매핑하도록 구성된 기계 학습 시스템을 포함한다. 음성들은 a) 제1 스피치 세그먼트로부터의 제1 복수의 분석적 오디오 세그먼트 및 b) 제2 스피치 세그먼트로부터의 제2 복수의 분석적 오디오 세그먼트의 주파수 분포의 함수로서 매핑된다.

[0015] 데이터베이스는 또한 제3 음성 내의 제3 스피치 세그먼트를 수신하도록 구성된다. 기계 학습 시스템은, 시간 수용 필드를 사용하여, 제3 스피치 세그먼트를 복수의 더 작은 분석적 오디오 세그먼트로 필터링하고, 벡터 공간에서 제1 음성 및 제2 음성에 대해 제3 음성을 매핑하도록 구성된다. 제1 음성 및 제2 음성에 대해 제3 음성을 매핑하는 것은 벡터 공간 내의 제2 음성에 대한 제1 음성의 상대적 위치를 변경한다.

[0016] 특히, 시스템은 적어도 하나의 음성에서 영어의 각각의 인간 음소(human phoneme)를 매핑하도록 구성된다. 수용 필드는 음성의 스피치 레이트 및/또는 액센트를 캡처하지 못하도록 충분히 작다. 예를 들어, 시간 수용 필드는 약 10 밀리초 내지 약 2000 밀리초 사이일 수 있다.

[0017] 다른 실시예에 따르면, 스피치 세그먼트들을 변환하기 위한 음색 벡터 공간을 구축하기 위한 방법은 a) 제1 음성 내의 제1 스피치 세그먼트 및 제2 음성 내의 제2 스피치 세그먼트를 수신하는 단계를 포함한다. 제1 스피치 세그먼트 및 제2 스피치 세그먼트 양자는 음색 데이터를 포함한다. 방법은 시간 수용 필드를 사용하여 제1 스피치 세그먼트 및 제2 스피치 세그먼트 각각을 복수의 더 작은 분석적 오디오 세그먼트로 필터링한다. 각각의 분석적 오디오 세그먼트는 음색 데이터를 나타내는 주파수 분포를 갖는다. 방법은 또한 제1 스피치 세그먼트 및 제2 스피치 세그먼트로부터의 복수의 분석적 오디오 세그먼트 중 적어도 하나에서의 주파수 분포의 함수로서 벡터 공간에서 제2 음성에 대해 제1 음성을 매핑한다.

[0018] 또한, 방법은 제3 음성 내의 제3 스피치 세그먼트를 수신하고, 시간 수용 필드를 사용하여 제3 스피치 세그먼트를 복수의 더 작은 분석적 오디오 세그먼트로 필터링할 수 있다. 제3 음성은 벡터 공간에서 제1 음성 및 제2 음성에 대해 매핑될 수 있다. 제1 음성의 제2 음성에 대한 상대적 위치는 제3 음성의 매핑의 함수로서 벡터 공간에서 조정될 수 있다.

[0019] 다른 실시예에 따르면, 음색 벡터 공간을 구축하기 위한 음색 벡터 공간 구축 시스템은 a) 제1 음성 내의 제1 음색 데이터를 포함하는 제1 스피치 세그먼트 및 b) 제2 음성 내의 제2 음색 데이터를 포함하는 제2 스피치 세그먼트를 수신하도록 구성된 입력을 포함한다. 시스템은 또한 a) 제1 스피치 세그먼트를 제1 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 갖는 제1 복수의 더 작은 분석적 오디오 세그먼트로, 그리고 b) 제2 스피치 세그먼트를 제2 복수의 더 작은 분석적 오디오 세그먼트로 필터링하기 위한 수단을 포함하고, 제2 복수의 더 작은 분석적 오디오 세그먼트 각각은 제2 음색 데이터의 상이한 부분을 나타내는 주파수 분포를 갖는다. 또한,

시스템은 a) 제1 스피치 세그먼트로부터의 제1 복수의 분석적 오디오 세그먼트 및 b) 제2 스피치 세그먼트로부터의 제2 복수의 분석적 오디오 세그먼트의 주파수 분포의 함수로서 음색 벡터 공간에서 제2 음성에 대해 제1 음성을 매핑하기 위한 수단을 갖는다.

[0020] 본 발명의 다른 실시예에 따르면, 음색 벡터 공간을 사용하여 새로운 음색을 갖는 새로운 음성을 구축하는 방법은 시간 수용 필드를 사용하여 필터링된 음색 데이터를 수신하는 단계를 포함한다. 음색 데이터는 음색 벡터 공간에서 매핑된다. 음색 데이터는 복수의 상이한 음성과 관련된다. 복수의 상이한 음성 각각은 음색 벡터 공간에서 각각의 음색 데이터를 갖는다. 방법은 기계 학습 시스템을 사용하여 복수의 상이한 음성의 음색 데이터를 사용하여 새로운 음색을 구축한다.

[0021] 일부 실시예들에서, 방법은 새로운 음성으로부터 새로운 스피치 세그먼트를 수신한다. 방법은 또한 신경망을 사용하여 새로운 스피치 세그먼트를 새로운 분석적 오디오 세그먼트로 필터링한다. 방법은 또한 복수의 매핑된 음성을 참조하여 벡터 공간에서 새로운 음성을 매핑한다. 방법은 또한, 새로운 음성과 복수의 매핑된 음성의 관계에 기초하여 새로운 음성의 적어도 하나의 특성을 결정한다. 특히, 특성은 성별, 인종 및/또는 나이일 수 있다. 복수의 음성 각각으로부터의 스피치 세그먼트는 상이한 스피치 세그먼트일 수 있다.

[0022] 일부 실시예들에서, 음색 데이터에 대한 수학적 연산의 함수로서, 후보 음성에서, 제1 후보 스피치 세그먼트를 생성하기 위해 생성적 신경망(generative neural network)이 사용된다. 예를 들어, 음색 데이터는 제1 음성 및 제2 음성에 관한 데이터를 포함할 수 있다. 또한, 벡터 공간 내의 음성 표현들의 클러스터(cluster of voice representations)는 특정 액센트를 나타낼 수 있다.

[0023] 일부 실시예들에서, 방법은 소스 스피치를 제공하고 소스 스피치를 새로운 음색으로 변환하면서 소스 케이던스(cadence) 및 소스 액센트를 유지한다. 시스템은 타겟 음색 데이터를 필터링하기 위한 수단을 포함할 수 있다.

[0024] 다른 실시예에 따르면, 시스템은 음색 벡터 공간을 사용하여 새로운 타겟 음성을 생성한다. 시스템은 시간 수용 필드를 사용하여 포함된 음색 데이터를 저장하도록 구성된 음색 벡터 공간을 포함한다. 음색 데이터는 시간 수용 필드를 사용하여 필터링된다. 음색 데이터는 복수의 상이한 음성과 관련된다. 기계 학습 시스템은 음색 데이터를 사용하여 음색 데이터를 새로운 타겟 음성으로 변환하도록 구성된다.

[0025] 많은 방식 중에서 특히, 음색 데이터는 음색 데이터의 적어도 하나의 음성 특성을 변수로서 사용하여 수학적 연산을 수행함으로써 새로운 타겟 음성으로 변환될 수 있다.

[0026] 또 다른 실시예에 따르면, 방법은 스피치 세그먼트를 소스 음색으로부터 타겟 음색으로 변환한다. 방법은 복수의 상이한 음성과 관련된 음색 데이터를 저장한다. 복수의 상이한 음성 각각은 음색 벡터 공간에서 각각의 음색 데이터를 갖는다. 음색 데이터는 시간 수용 필드를 사용하여 필터링되고 음색 벡터 공간에서 매핑된다. 방법은 타겟 음성으로 변환하기 위한 소스 음성 내의 소스 스피치 세그먼트를 수신한다. 방법은 또한 타겟 음성의 선택을 수신한다. 타겟 음성은 타겟 음색을 갖는다. 타겟 음성은 복수의 상이한 음성을 참조하여 음색 벡터 공간에서 매핑된다. 방법은 기계 학습 시스템을 사용하여 소스 스피치 세그먼트를 소스 음성의 음색으로부터 타겟 음성의 음색으로 변환한다.

[0027] 본 발명의 예시적인 실시예들은 컴퓨터 판독가능 프로그램 코드를 갖는 컴퓨터 사용가능 매체를 갖는 컴퓨터 프로그램 제품으로서 구현된다.

도면의 간단한 설명

[0028] 본 기술분야의 통상의 기술자들은 바로 아래에 요약된 도면들을 참조하여 논의되는 아래의 "예시적 실시예들의 설명"으로부터 본 발명의 다양한 실시예들의 이점들을 더 충분히 이해해야 한다.

도 1은 본 발명의 예시적인 실시예들에 따른 음성 대 음성 변환 시스템의 단순화된 버전을 개략적으로 도시한다.

도 2는 본 발명의 예시적인 실시예들을 구현하는 시스템의 상세들을 개략적으로 도시한다.

도 3은 본 발명의 예시적인 실시예들에 따른, 인코딩된 음성 데이터를 나타내는 다차원 공간을 구축하기 위한 프로세스를 도시한다.

도 4는 본 발명의 예시적인 실시예들에 따른 스피치 샘플을 필터링하는 시간 수용 필터를 개략적으로 도시한다.

도 5a-5c는 본 발명의 예시적인 실시예들에 따른, 도 4의 동일한 스피치 세그먼트로부터의 상이한 분석적 오디

오 세그먼트들의 추출된 주파수 분포들을 갖는 스펙트로그램들(spectrograms)을 도시한다.

도 5a는 "Call"이라는 단어 내의 "a"라는 단음에 대한 스펙트로그램을 도시한다.

도 5b는 "Stella" 내의 "a"라는 단음에 대한 스펙트로그램을 도시한다.

도 5c는 "Please" 내의 "ea"라는 단음에 대한 스펙트로그램을 도시한다.

도 6a-6d는 본 발명의 예시적인 실시예들에 따른 벡터 공간의 슬라이스들(slices)을 개략적으로 도시한다.

도 6a는 도 5b에 도시된 단음에 대한 타겟 음성만을 매핑하는 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 6b는 타겟 음성 및 제2 음성을 매핑하는 도 6a의 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 6c는 타겟 음성, 제2 음성 및 제3 음성을 매핑하는 도 6a의 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 6d는 복수의 음성을 매핑하는 도 6a의 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 7a는 제2 음성의 음색 내의 "Call"이라는 단어 내의 "a"라는 단음에 대한 스펙트로그램을 도시한다.

도 7b는 제3 음성의 음색 내의 "Call"이라는 단어 내의 "a"라는 단음에 대한 스펙트로그램을 도시한다.

도 8a는 본 발명의 예시적인 실시예들에 따른 합성 음성 프로파일을 포함하는 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 8b는 본 발명의 예시적인 실시예들에 따른, 생성적 적대적 신경망이 합성 음성 프로파일을 세밀화(refine)한 후에 "DOG" 내의 "D"라는 단음에 대응하는 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 8c는 제2 음성 및 제4 음성의 추가를 갖는 도 8b의 벡터 공간의 슬라이스를 개략적으로 도시한다.

도 9는 본 발명의 예시적인 실시예들에 따른 증강된 음성 프로파일을 세밀화하기 위해 생성적 적대적 망을 사용하는 시스템의 블록도를 도시한다.

도 10은 본 발명의 예시적인 실시예들에 따른 스피치 대 스피치(speech-to-speech) 변환 프로세스를 도시한다.

도 11은 본 발명의 예시적인 실시예들에 따른 음성을 사용하여 아이덴티티를 검증하는 프로세스를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0029] 예시적인 실시예들에서, 음성 대 음성 변환 시스템은 소스 음성에서 발음된 스피치 세그먼트의 타겟 음성으로의 실시간 또는 거의 실시간 변환을 가능하게 한다. 이를 위해, 시스템은 복수의 음성으로부터 스피치 샘플들을 수신하고 각각의 음성에 의해 만들어진 각각의 사운드와 연관된 주파수 성분들을 추출하는 음성 특징 추출기를 갖는다. 음성들은 추출된 주파수 성분들에 기초하여 서로에 대해 벡터 공간에서 매핑되며, 이는 스피치 샘플들에서 제공되지 않는 사운드들에 대한 합성 주파수 성분들의 추정(extrapolation)을 가능하게 한다. 시스템은 타겟 음성을 다른 음성들과 비교하고, 합성 주파수 성분들을 세밀화하여 음성을 최적으로 모방하도록 더 구성된 기계 학습을 갖는다. 따라서, 시스템의 사용자들은 스피치 세그먼트를 입력하고, 타겟 음성을 선택할 수 있고, 시스템은 스피치 세그먼트를 타겟 음성으로 변환한다.

[0030] 도 1은 본 발명의 예시적인 실시예들에 따른 음성 대 음성 변환 시스템(100)의 단순화된 버전을 개략적으로 도시한다. 특히, 시스템(100)은 사용자가 자신의 음성(또는 임의의 다른 음성)을 그가 선택한 타겟 음성(104)으로 변환하는 것을 가능하게 한다. 보다 구체적으로, 시스템(100)은 사용자의 스피치 세그먼트(103)를 타겟 음성(104)으로 변환한다. 따라서, 이 예에서 사용자의 음성은 소스 음성(102)으로 지칭되는데, 그 이유는 시스템(100)이 소스 음성(102)에서 발음된 스피치 세그먼트(103)를 타겟 음성(104)으로 변환하기 때문이다. 변환의 결과는 변환된 스피치 세그먼트(106)이다. 소스 음성(102)이 인간 화자(예를 들어, 아놀드)로서 도시되지만, 일부 실시예들에서 소스 음성(102)은 합성 음성일 수 있다.

[0031] 음성들의 변환은 음색 변환으로도 지칭된다. 본 출원 전체에 걸쳐, "음성(voice)" 및 "음색(timbre)"은 상호 교환 가능하게 사용된다. 음성들의 음색은 청취자들이 동일한 피치, 액센트, 진폭 및 케이던스에서 동일한 단어들을 달리 발음하고 있는 특정 음성들을 구별하고 식별할 수 있게 한다. 음색은 화자가 특정 사운드를 위해 만들어내는 주파수 성분들의 세트로부터 초래되는 생리학적 특성(physiological property)이다. 예시적인 실시예들에서, 스피치 세그먼트(103)의 음색은 소스 음성(102)의 원래 케이던스, 리듬 및 악센트/발음을 유지하면서 타겟 음성(104)의 음색으로 변환된다.

[0032] 일례로서, 아놀드 슈왈제네거는 시스템(100)을 사용하여 그의 스피치 세그먼트(103)(예를 들어, "I'll be back")를 제임스 얼 존스의 음성/음색으로 변환할 수 있다. 이 예에서, 아놀드의 음성은 소스 음성(102)이고, 제임스의 음성은 타겟 음성(104)이다. 아놀드는 제임스의 음성의 스피치 샘플(105)을 시스템(100)에 제공할 수 있으며, 시스템은 스피치 샘플(105)을 사용하여 그의 스피치 세그먼트를 변환한다(아래에 더 설명됨). 시스템(100)은 스피치 세그먼트(103)를 취하고, 그것을 제임스의 음성(104)으로 변환하고, 변환된 스피치 세그먼트(106)를 타겟 음성(104)에서 출력한다. 따라서, 스피치 세그먼트(103) "I'll be back"은 제임스의 음성(104)에서 출력된다. 그러나, 변환된 스피치 세그먼트(106)는 원래 케이던스, 리듬 및 액센트를 유지한다. 따라서, 변환된 스피치 세그먼트(106)는 제임스가 아놀드의 액센트/발음/케이던스 및 스피치 세그먼트(103)를 모방하려고 시도하고 있는 것처럼 들린다. 즉, 변환된 스피치 세그먼트(106)는 제임스의 음색에서의 소스 스피치 세그먼트(103)이다. 시스템(100)이 이러한 변환을 어떻게 달성하는지에 대한 상세들이 이하에서 설명된다.

[0033] 도 2는 본 발명의 예시적인 실시예들을 구현하는 시스템(100)의 상세들을 개략적으로 도시한다. 시스템(100)은 오디오 파일들, 예를 들어 타겟 음성(104) 내의 스피치 샘플(105) 및 소스 음성(102)으로부터의 스피치 세그먼트들(103)을 수신하도록 구성된 입력(108)을 갖는다. 상이한 용어들이 "스피치 세그먼트(103)" 및 "스피치 샘플(105)"에 대해 사용되지만, 이들 양자는 발음된 단어들(spoken words)을 포함할 수 있다는 것을 이해해야 한다. "스피치 샘플(105)" 및 "스피치 세그먼트(103)"라는 용어는 단지 소스를 나타내는 데 사용되며, 시스템(100)은 이들 오디오 파일 각각에 대해 상이한 변환들을 행한다. "스피치 샘플(105)"은 타겟 음성(104)에서 시스템(100)에 입력되는 스피치를 말한다. 시스템(100)은 스피치 샘플(105)을 사용하여 타겟 음성(104)의 주파수 성분들을 추출한다. 한편, 시스템(100)은 소스 음성(102)으로부터의 "스피치 세그먼트(103)"를 타겟 음성(104)으로 변환한다.

[0034] 시스템(100)은 사용자가 시스템(100)과 통신할 수 있는 사용자 인터페이스를 제공하도록 구성된 사용자 인터페이스 서버(110)를 갖는다. 사용자는 전자 디바이스(예컨대, 컴퓨터, 스마트폰 등)를 통해 사용자 인터페이스에 액세스하고, 전자 디바이스를 사용하여 스피치 세그먼트(103)를 입력(108)에 제공할 수 있다. 일부 실시예들에서, 전자 디바이스는 인터넷 접속된 스마트폰 또는 데스크톱 컴퓨터와 같은 네트워킹된 디바이스일 수 있다. 사용자 스피치 세그먼트(103)는 예를 들어 사용자에 의해 발음된 문장(예를 들어, "I'll be back")일 수 있다. 이를 위해, 사용자 디바이스는 사용자 스피치 세그먼트(103)를 기록하기 위한 통합 마이크로폰 (integrated microphone)또는 보조 마이크로폰(예를 들어, USB에 의해 접속됨)을 가질 수 있다. 대안적으로, 사용자는 사용자 스피치 세그먼트(103)를 포함하는 미리 기록된 디지털 파일(예를 들어, 오디오 파일)을 업로드할 수 있다. 사용자 스피치 세그먼트(103) 내의 음성은 반드시 사용자의 음성일 필요는 없다는 것을 이해해야 한다. 용어 "사용자 스피치 세그먼트(103)"는 편의상 시스템(100)이 타겟 음색으로 변환하는 사용자에 의해 제공된 스피치 세그먼트를 나타내기 위해 사용된다. 전술한 바와 같이, 사용자 스피치 세그먼트(103)는 소스 음성(102)에서 발음된다.

[0035] 입력(108)은 또한 타겟 음성(104)을 수신하도록 구성된다. 이를 위해, 스피치 세그먼트(103)와 유사한 방식으로, 타겟 음성(104)이 사용자에 의해 시스템(100)에 업로드될 수 있다. 대안적으로, 타겟 음성(104)은 시스템(100)에 이전에 제공된 음성들(111)의 데이터베이스에 있을 수 있다. 이하에서 더 상세히 설명되는 바와 같이, 타겟 음성(104)이 음성들(111)의 데이터베이스에 이미 있지 않은 경우, 시스템(100)은 변환 엔진(118)을 사용하여 음성(104)을 처리하고 이를 인코딩된 음성 데이터를 나타내는 다차원 이산 또는 연속 공간(112)에서 매핑한다. 표현(representation)은 음성들을 "매핑(mapping)"하는 것으로 지칭된다. 인코딩된 음성 데이터가 매핑될 때, 벡터 공간(112)은 음성들에 관한 특성화들(characterizations)을 행하고, 이들을 그에 기초하여 서로에 대해 배치한다. 예를 들어, 표현의 일부는 음성의 피치 또는 화자의 성별과 관련이 있을 수 있다.

[0036] 예시적인 실시예들은 시간 수용 필터(114)(시간 수용 필드(114)라고도 함)를 사용하여 타겟 음성(104)을 분석적 오디오 세그먼트들로 필터링하고, 변환 엔진(118)은 분석적 오디오 세그먼트들로부터 주파수 성분들을 추출하고, 기계 학습 시스템(116)은 타겟 음성(104)이 입력(108)에 의해 처음 수신될 때 (예를 들어, 음성 특징 추출기(120)를 사용하여) 벡터 공간(112)에서 타겟 음성(104)의 표현을 매핑하고, 기계 학습 시스템(116)은 타겟 음성(104)의 매핑된 표현을 세밀화한다. 이어서, 시스템(100)은 스피치 세그먼트들(103)을 타겟 음성(104)으로 변환하는 데 사용될 수 있다.

[0037] 특히, 예시적인 실시예들에서, 시스템(100)은 타겟(104) 스피치 샘플(105)을 (잠재적으로 중첩되는) 오디오 세그먼트들로 분할하고, 이들 각각은 음성 특징 추출기(120)의 시간 수용 필드(114)에 대응하는 크기를 갖는다. 이어서, 음성 특징 추출기(120)는 각각의 분석적 오디오 세그먼트에 대해 개별적으로 작용하며, 이들 각각은 타겟 화자의 음성(104)에서 타겟에 의해 만들어진 사운드(예를 들어, 단음, 음소, 단음의 일부 또는 다수의 단

음)를 포함할 수 있다.

[0038] 각각의 분석적 오디오 세그먼트에서, 음성 특징 추출기(120)는 타겟 화자의 음성(104)의 특징들을 추출하고, 그러한 특징들에 기초하여 벡터 공간(112)에서 음성들을 매핑한다. 예를 들어, 하나의 이러한 특징은 일부 모음 사운드들을 생성하기 위해 사용된 몇몇 주파수의 일부 진폭들을 증폭하는데 편향될 수 있고, 추출 방법은 세그먼트 내의 사운드를 특정 모음 사운드로서 식별하고, 표현된 주파수들의 진폭들을 다른 음성들에 의해 사용된 것들과 비교하여 유사한 사운드들을 생성하고, 이어서 음성 특징 추출기(120)가 이전에 특징으로서 노출되었던 유사한 음성들의 특정 세트와 비교하여 이 음성의 주파수들의 차이를 인코딩할 수 있다. 이어서, 이들 특징은 함께 결합되어 타겟 음성(104)의 매핑된 표현을 세밀화한다.

[0039] 예시적인 실시예들에서, 시스템(100)(말단부에서의 결합과 함께 음성 특징 추출기(120))은 기계 학습 시스템으로 간주될 수 있다. 일 구현은 음성 특징 추출기(120)로서의 컨볼루션 신경망(convolutional neural network), 및 말단부에서 추출된 특징들을 결합하기 위한 순환 신경망(recurrent neural network)을 포함할 수 있다. 다른 예들은 말단부에서의 주의 메커니즘(attention mechanism)을 갖는 신경망 또는 말단부에서의 고정 크기 신경망(fixed-sized neural network) 또는 말단부에서의 특징들의 간단한 추가와 함께 컨볼루션 신경망을 포함할 수 있다.

[0040] 음성 특징 추출기(120)는 타겟 스피치 샘플(105)의 주파수들에서의 진폭들(예를 들어, 포먼트들(formants)의 상대 진폭들(relative amplitudes) 및/또는 포맷들의 공격 및 감쇠(attack and decay)) 간의 관계들을 추출한다. 그렇게 함으로써, 시스템(100)은 타겟의 음색(104)을 학습하고 있다. 일부 실시예들에서, 음성 특징 추출기(120)는 선택적으로 특정 분석적 오디오 세그먼트 내의 주파수 성분들을 특정 사운드와 상관시키는 주파수 대 사운드 상관 엔진(frequency-to-sound correlation engine, 122)을 포함할 수 있다. 주파수 대 사운드 상관 엔진(122)이 타겟 음성(104)을 매핑하기 위해 사용되는 것으로 위에서 설명되었지만, 본 기술분야의 통상의 기술자는 기계 학습 시스템(116)이 음성들을 매핑하기 위해 추가적인 또는 대안적인 방법들을 사용할 수 있다는 것을 이해한다. 따라서, 이러한 특정 구현의 논의는 단지 논의를 용이하게 하기 위한 예로서 의도된 것이고, 모든 예시적인 실시예들을 제한하려는 의도는 아니다.

[0041] 전술한 컴포넌트들 각각은 임의의 종래의 상호접속 메커니즘에 의해 동작 가능하게 접속된다. 도 2는 컴포넌트들 각각과 통신하는 버스를 간단히 도시한다. 본 기술분야의 통상의 기술자들은 이러한 일반화된 표현이 다른 종래의 직접 또는 간접 접속들을 포함하도록 수정될 수 있다는 것을 이해해야 한다. 따라서, 버스에 대한 논의는 다양한 실시예들을 제한하는 것으로 의도되지 않는다.

[0042] 실제로, 도 2는 이러한 컴포넌트들 각각을 단지 개략적으로 도시한다는 점에 유의해야 한다. 본 기술분야의 통상의 기술자들은 이러한 컴포넌트들 각각이 하나 이상의 다른 기능적 컴포넌트들에 걸쳐 하드웨어, 소프트웨어, 또는 하드웨어와 소프트웨어의 조합을 사용하는 것 등에 의해 다양한 종래의 방식들로 구현될 수 있다는 것을 이해해야 한다. 예를 들어, 음성 추출기(112)는 펌웨어를 실행하는 복수의 마이크로프로세서를 사용하여 구현될 수 있다. 다른 예로서, 기계 학습 시스템(116)은 하나 이상의 주문형 집적회로(즉, "ASIC") 및 관련 소프트웨어, 또는 ASIC들, 개별 전자 컴포넌트들(예를 들어, 트랜지스터들) 및 마이크로프로세서들의 조합을 사용하여 구현될 수 있다. 따라서, 도 2의 단일 박스 내의 기계 학습 시스템(116) 및 다른 컴포넌트들의 표현은 단지 간소화를 위한 것이다. 사실, 일부 실시예들에서, 도 2의 기계 학습 시스템(116)은, 반드시 동일한 하우징 또는 섀시 내에 있지는 않는 복수의 상이한 기계에 걸쳐 분포된다. 또한, 일부 실시예들에서, 개별적인 것으로서 도시된 컴포넌트들(예를 들어, 도 2의 시간 수용 필드들(114))은 단일 컴포넌트(예를 들어, 전체 기계 학습 시스템(116)에 대한 단일 시간 수용 필드(115))로 대체될 수 있다. 또한, 도 2의 소정 컴포넌트들 및 서브컴포넌트들은 선택적이다. 예를 들어, 일부 실시예들은 상관 엔진을 사용하지 않을 수 있다. 다른 예로서, 일부 실시예들에서, 생성기(140), 판별기(142) 및/또는 음성 특징 추출기(120)는 수용 필드(114)를 갖지 않을 수 있다.

[0043] 도 2의 표현은 실제 음성 대 음성 변환 시스템(100)의 상당히 단순화된 표현이라는 것을 반복해야 한다. 이 분야의 기술자들은 이러한 디바이스가 중앙 처리 유닛들, 다른 패킷 처리 모듈들 및 단기 메모리와 같은 다른 물리적 및 기능적 컴포넌트들을 가질 수 있다는 것을 이해해야 한다. 따라서, 이 논의는 도 2가 음성 대 음성 변환 시스템(100)의 모든 요소들을 나타낸다는 것을 암시하도록 의도되지 않는다.

[0044] 도 3은 본 발명의 예시적인 실시예들에 따른, 인코딩된 음성 데이터를 나타내는 다차원 이산 또는 연속 벡터 공간(112)을 구축하기 위한 프로세스(300)를 도시한다. 이 프로세스는 벡터 공간(112)을 구축하기 위해 통상적으로 사용될 더 긴 프로세스로부터 실질적으로 단순화된다는 점에 유의해야 한다. 따라서, 벡터 공간(112)을 구축하는 프로세스는 이 분야의 기술자들이 사용할 가능성이 있는 많은 단계를 가질 수 있다. 또한, 단계들 중

일부는 도시된 것과 상이한 순서로 또는 동시에 수행될 수 있다. 따라서, 이 분야의 통상의 기술자들은 프로세스를 적절하게 수정할 수 있다.

[0045] 도 3의 프로세스는 단계 302에서 시작하며, 이 단계에서 타겟 음색(104)에 있는 스피치 샘플(105)을 수신한다. 전술한 바와 같이, 스피치 샘플(105)은 입력(108)에 의해 수신되고, 시스템(100)의 사용자에 의해 시스템(100)에 제공될 수 있다. 일부 실시예들에서, 시스템(100)은 벡터 공간(112)에서 이미 매핑된 음성들을 제공받을 수 있다. 벡터 공간(112)에서 이미 매핑된 음성들은 이하에서 설명되는 프로세스를 이미 겪었다. 벡터 공간(112)은 이하에서 더 상세히 설명된다.

[0046] 도 4는 본 발명의 예시적인 실시예들에 따른 스피치 샘플(105)을 필터링하는 예시적인 시간 수용 필터(114)를 개략적으로 도시한다. 프로세스는 단계 304로 계속되고, 여기서 스피치 샘플(105)은 시간 수용 필터(114)에 의해 분석적 오디오 세그먼트들(124)로 필터링된다. 이 예에서의 스피치 샘플(105)은 타겟 음성(104)에서의 1초 기록된 오디오 신호이다. 스피치 샘플(105)은 1초보다 짧거나 길 수 있지만, 이하에서 논의되는 이유들로 인해, 일부 실시예들은 스피치 샘플(105)에 대해 더 긴 길이를 사용할 수 있다. 이 예에서의 시간 수용 필터(114)는 100 밀리초로 설정된다. 따라서, 1초 스피치 샘플(105)은 필터(114)에 의해 10개의 100 밀리초 분석적 오디오 세그먼트(124)로 분할된다.

[0047] 시간 수용 필터(114)가 100 밀리초 간격들을 필터링하도록 설정되는 것으로 도시되지만, 다양한 필터링 간격들이 이하에서 논의되는 바와 같은 파라미터들 내에서 설정될 수 있다는 것을 이해해야 한다. 시간 수용 필드(114)(또는 필터(114))의 논의는 기계 학습(116)(예를 들어, 생성기(140), 판별기(142) 및/또는 특징 추출기(120))의 임의의 또는 모든 부분들에 관련된다. 예시적인 실시예들에서, 필터링 간격은 0 밀리초보다 크고 300 밀리초보다 작다. 일부 다른 실시예들에서, 시간 수용 필드(114)는 50 밀리초, 80 밀리초, 100 밀리초, 150 밀리초, 250 밀리초, 400 밀리초, 500 밀리초, 600 밀리초, 700 밀리초, 800 밀리초, 900 밀리초, 1000 밀리초, 1500 밀리초 또는 2000 밀리초 미만이다. 추가 실시예들에서, 시간 수용 필드(114)는 5 밀리초, 10 밀리초, 15 밀리초, 20 밀리초, 30 밀리초, 40 밀리초, 50 밀리초, 또는 60 밀리초보다 크다. 도 2에서 개별 컴포넌트로서 도시되지만, 시간 수용 필터(114)는 시간 수용 필드(114)로서 입력(108)에 내장될 수 있다. 또한, 기계 학습 시스템(116)은 (예를 들어, 도시된 3개의 개별 수용 필드(114) 대신에) 단일 수용 필드(114)를 가질 수 있다.

[0048] 각각의 분석적 오디오 세그먼트(124)는 특정 타겟 음성(104)에 의해 만들어진 특정 사운드 또는 사운드들에 대한 주파수 데이터(단계 306에서 추출됨)를 포함한다. 따라서, 분석적 오디오 세그먼트(124)가 더 짧을수록, 주파수 데이터(예를 들어, 주파수들의 분포)는 특정 사운드에 더 구체적이다. 그러나, 분석적 오디오 세그먼트(124)가 너무 짧으면, 소정 저주파 사운드들은 시스템(100)에 의해 필터링될 수 있는 것이 가능하다. 바람직한 실시예들에서, 시간 필터(114)는 스피치 샘플(105)의 스트림 내의 사운드의 최소 구별 가능 이산 세그먼트(the smallest distinguishable discrete segment)를 캡처하도록 설정된다. 사운드의 최소 구별 가능 이산 세그먼트는 단음으로 지칭된다. 기술적 관점에서, 분석적 오디오 세그먼트(124)는 단음의 포먼트 특성들을 캡처하기에 충분히 짧아야 한다. 예시적인 실시예들은 분석적 오디오 세그먼트들을 약 60 밀리초 내지 약 250 밀리초 사이로 필터링할 수 있다.

[0049] 사람들은 일반적으로 20 Hz 내지 20 kHz 범위 내의 사운드들을 청취할 수 있다. 더 낮은 주파수의 사운드들은 더 높은 주파수의 사운드들보다 더 긴 주기(period)를 갖는다. 예를 들어, 20 Hz 주파수를 갖는 사운드 파(sound wave)는 전체 주기 동안 50 밀리초 걸리는 반면, 2 kHz 주파수를 갖는 사운드 파는 전체 주기 동안 0.5 밀리초 걸린다. 따라서, 분석적 오디오 세그먼트(124)가 매우 짧은 경우(예를 들어, 1 밀리초), 분석적 오디오 세그먼트(124)는 검출 가능하기에 충분한 20 Hz 사운드를 포함하지 않을 수 있는 것이 가능하다. 그러나, 일부 실시예들은 예측 모델링(predictive modeling)을 사용하여(예를 들어, 저주파 사운드 파의 일부만을 사용하여) 더 낮은 주파수의 사운드들을 검출할 수 있다. 예시적인 실시예들은 일부 저주파 사운드들을 필터링 또는 무시하고, 타겟 음성(104)의 음색을 정확하게 모방하기에 충분한 주파수 데이터를 여전히 포함할 수 있다. 따라서, 본 발명자들은 약 10 밀리초 정도로 짧은 분석적 오디오 세그먼트들(124)이 시스템(100)으로 하여금 단음들의 주파수 특성들을 적절히 예측하게 하기에 충분하다고 생각한다.

[0050] 사람의 음성에서의 기본 주파수(fundamental frequency)는 일반적으로 100 Hz보다 큰 정도이다. 기본 주파수는 음색의 일부이지만, 음색 자체는 아니다. 사람의 음성들이 그들의 기본 주파수에서만 상이한 경우, 음성 변환은 본질적으로 피치-시프팅, 즉 동일한 노래를 피아노로 한 옥타브 낮게 연주하는 것의 등가물일 것이다. 그러나, 음색은 또한 동일한 음표를 연주하는 피아노 및 트럼펫 사운드를 상이하게 만드는 품질이고- 주파수에서의 모든 작은 추가적인 변화들(variations)의 집합이며, 이들 중 어느 것도 (일반적으로) 기본 주파수만큼 높은 진

폭을 갖지 않지만, 사운드의 전체 느낌에 상당히 기여한다.

[0051] 기본 주파수는 음색에 중요할 수 있지만, 그것은 단독으로는 음색의 유일한 표시자(indicator)가 아니다. 모건 프리먼 및 타겟 음성(104) 둘 다가 동일한 옥타브에서 동일한 음표들 중 일부를 칠 수 있는 경우를 고려한다. 이러한 음표들은 동일한 기본 주파수를 암시적으로 갖지만, 타겟 음성(104) 및 모건 프리먼은 상이한 음색들을 가질 수 있고, 따라서 기본 주파수 단독으로는 음성을 식별하기에 충분하지 않다.

[0052] 시스템(100)은 궁극적으로 분석적 오디오 세그먼트들(124)로부터의 주파수 데이터에 기초하여 타겟 음성(104)에 대한 음성 프로파일을 생성한다. 따라서, 특정 단음에 대응하는 주파수 데이터를 갖기 위해, 시간 수용 필터(114)는 바람직하게 분석적 오디오 세그먼트들(124)을 최소 구별 가능 단음을 발음하는 데 걸리는 시간으로 대략 필터링한다. 상이한 단음들은 상이한 시간 길이들(즉, 단음을 발음하는 데 걸리는 시간의 양)을 가질 수 있기 때문에, 예시적인 실시예들은 분석적 오디오 세그먼트들(124)을 인간 언어들에서 만들어지는 가장 긴 단음을 발음하는 데 걸리는 시간보다 큰 길이로 필터링할 수 있다. 예시적인 실시예들에서, 필터(114)에 의해 설정된 시간 플로어(temporal floor)는 분석적 오디오 세그먼트(124)가 단일 단음의 적어도 전체에 관한 주파수 정보를 포함하는 것을 가능하게 한다. 본 발명자들은 스피치를 100 밀리초 분석적 오디오 세그먼트들(124)로 분할하는 것이 인간 음성들에 의해 만들어지는 대부분의 단음들에 대응할 정도로 충분히 짧다고 생각한다. 따라서, 각각의 분석적 오디오 세그먼트(124)는 스피치 샘플(105)에서 타겟 음성(104)에 의해 만들어진 소정의 사운드들(예를 들어, 단음들)에 대응하는 주파수 분포 정보를 포함한다.

[0053] 한편, 예시적인 실시예들은 또한 시간 수용 필드(114)에 대한 상한(ceiling)을 가질 수 있다. 예를 들어, 예시적인 실시예들은 한 번에 하나보다 많은 완전한 단음을 캡처하는 것을 회피하기에 충분히 짧은 수용 필드(114)를 갖는다. 또한, 시간 수용 필드(114)가 큰 경우(예를 들어, 1초보다 큰 경우), 분석적 오디오 세그먼트들(124)은 소스(102)의 액센트 및/또는 케이던스를 포함할 수 있다. 일부 실시예들에서, 시간 수용 필드(114)는 액센트 또는 케이던스 음성 특성들의 캡처를 회피하기에 충분히 짧다(즉, 상한을 갖는다). 이러한 음성 특성들은 더 긴 시간 간격들에 걸쳐 픽업된다.

[0054] 일부 종래 기술의 텍스트 대 스피치 변환 시스템들은 액센트를 포함한다. 예를 들어, 미국 액센트는 "zebra"라는 단어를 ['zi:brə]("zeebrah")로 발음할 수 있고, 영국 액센트는 그 단어를 ['zɛbrə]("zebrah")로 발음할 수 있다. 미국 및 영국 화자들 양자는 상이한 단어들에서 i: 및 ɛ 단음들 양자를 사용하지만, 텍스트 대 스피치는 액센트에 기초하여 특정 단어 "zebra"에서 하나의 단음 또는 다른 하나의 단음을 사용한다. 따라서, 텍스트 대 스피치는 타겟 음색의 완전한 제어를 허용하지 않는 대신에 타겟이 특정 단어들을 발음하는 방식에 의해 제한된다. 따라서, 충분히 짧은 수용 필드(114)를 유지함으로써, 분석적 오디오 세그먼트들(124)은 주로 (예를 들어, 완전한 단어 "zebra"에서) 더 긴 시간 간격들에 걸쳐 픽업된 이러한 다른 특성들을 포함하는 데이터를 수집하는 것을 회피한다.

[0055] 사실상, 본 발명자들에게 알려진 종래 기술은 수용 필드들이 너무 길기 때문에 순수한 음색을 캡처하는 문제들을 갖는데, 예를 들어 수용 필드들은 음성 매핑으로 하여금 음색(예를 들어, 액센트)을 매핑하려고 시도할 때 추가적인 특성들을 본질적으로 포함하게 한다. 액센트 매핑의 문제는 화자가 화자의 음색을 유지하면서 액센트를 변경할 수 있다는 것이다. 따라서, 그러한 종래 기술은 이러한 다른 특성들로부터 분리된 음성의 진정한 음색을 획득할 수 없다. 예를 들어, Arik et al. (Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou: Neural Voice Cloning with a Few Samples., arXiv:1708.07524, 2018)에 설명된 것과 같은 종래 기술의 텍스트 대 스피치 변환은 변환된 단어에 기초하여 전체 음성을 합성한다. 변환은 스피치 대 스피치가 아니라 텍스트 대 스피치이기 때문에, 시스템은 음색에 대해서뿐만 아니라, 케이던스, 굴절(inflection), 액센트 등에 대해서도 결정을 행해야 한다. 대부분의 텍스트 대 스피치 시스템들은 이러한 특성들 각각을 개별적으로 결정하지 않는 대신에, 그들이 트레이닝되는 각각의 사람에 대해, 그 사람에 대한 이들 요소 모두의 조합을 학습한다. 이것은 음색에 대한 음성의 개별적인 조정이 없다는 것을 의미한다.

[0056] 대조적으로, 예시적인 실시예들은 스피치를 합성하는 것이 아니라, 스피치 대 스피치 변환(음성 대 음성 변환이라고도 함)을 사용하여 스피치를 변환한다. 시스템(100)은 케이던스, 액센트 등과 같은 모든 다른 특성들에 대해 선택할 필요가 없는데, 이는 이러한 특성들이 입력 스피치에 의해 제공되기 때문이다. 따라서, 입력 스피치(예를 들어, 스피치 세그먼트(103))는 다른 스피치 특성들을 유지하면서 상이한 음색으로 구체적으로 변환된다.

[0057] 도 3으로 돌아가서, 프로세스는 단계 306으로 진행하여, 분석적 오디오 세그먼트들(124)로부터 주파수 분포들을 추출한다. 임의의 특정 분석적 오디오 세그먼트(124)의 주파수 분포는 음성마다 상이하다. 이것은 상이한 화

자들의 음색들이 구별 가능한 이유이다. 특정 분석적 오디오 세그먼트(124)로부터 주파수 정보를 추출하기 위해, 변환 엔진(118)은 단시간 푸리에 변환(STFT)을 수행할 수 있다. 그러나, STFT는 주파수 데이터를 획득하는 하나의 방식일 뿐이라는 것을 이해해야 한다. 예시적인 실시예들에서, 변환 엔진(118)은 기계 학습의 일부일 수 있고, 또한 주파수 데이터를 생성하는 그 자신의 필터들의 세트를 구축할 수 있다. 스피치 샘플(105)은 (잠재적으로 중첩되는) 분석적 오디오 세그먼트들(124)로 분할되고, 변환 엔진은 각각의 분석적 오디오 세그먼트(124)에 대해 FFT를 수행한다. 일부 실시예들에서, 변환 엔진(118)은 경계 조건들(boundary conditions)을 갖는 문제들을 완화하기 위해 분석적 오디오 세그먼트(124)에 대한 윈도우잉 기능(windowing function)을 포함한다. 분석적 오디오 세그먼트들(124) 사이에 소정의 중첩이 존재하더라도, 이들은 여전히 상이한 오디오 세그먼트들(124)로 간주된다. 추출이 완료된 후에, 분석적 오디오 세그먼트들(124) 주파수 데이터가 획득된다. 그 결과는 다양한 시점들에서의 주파수 강도들의 세트이며, 이들은 예시적인 실시예들에서 수직축 상의 주파수 및 수평축 상의 시간을 갖는 이미지(스펙트로그램)로서 배열된다.

[0058] 도 5a-5c는 본 발명의 예시적인 실시예들에 따른, 도 4의 동일한 스피치 샘플(105)로부터의 상이한 분석적 오디오 세그먼트들(124)의 추출된 주파수 분포들을 갖는 스펙트로그램들(126)을 도시한다. "주파수 분포들"이라는 용어는 상황에 따라 특정 분석적 오디오 세그먼트(124) 또는 그들의 집합에 존재하는 개별적인 주파수들 및 그들의 개별적인 강도들의 세트를 지칭한다. 도 5a는 타겟(104)에 의해 만들어진 단어 "Call" 내의 "a"라는 단음에 대한 스펙트로그램(126)을 도시한다. 이 분야의 기술자들에게 알려진 바와 같이, 스펙트로그램(126)은 주파수에 대한 시간을 플로팅하고, 또한 컬러 강도를 통해 주파수의 진폭/강도(예를 들어, dB 단위)를 나타낸다. 도 5a에서, 스펙트로그램(126)은 (포먼트들(128)로도 지칭되는) 12개의 명확하게 가시적인 피크(128)를 가지며, 각각의 피크는 주파수가 더 잘 들리는 것과 연관된 컬러 강도를 갖는다.

[0059] 시스템(100)은 도 5a의 스펙트로그램이 "a" 사운드를 나타낸다는 것을 안다. 예를 들어, 상관 엔진(122)은 분석적 오디오 세그먼트들(124)에 대한 주파수 분포를 분석할 수 있고, 이 주파수 분포가 단어 "Call" 내의 "a" 단음을 나타내는 것으로 결정할 수 있다. 시스템(100)은 단음을 결정하기 위해 분석적 오디오 세그먼트(124)의 주파수 성분들을 사용한다. 예를 들어, "Call" 내의 "a" 사운드는 누가 말하고 있는지에 관계없이 중간 주파수 성분들(2 kHz 근처)을 갖는 반면, 그러한 주파수 성분들은 다른 모음 사운드들에 대해서는 존재하지 않을 수 있다. 시스템(100)은 사운드를 추측하기 위해 주파수 성분들의 차이들(distinctions)을 사용한다. 또한, 시스템(100)은 이 주파수 분포 및 강도가 타겟(104)에 특유하다는 것을 안다. 타겟(104)이 동일한 "a" 단음을 반복하면, 동일하지는 않더라도 매우 유사한 주파수 분포가 존재한다.

[0060] 특징 추출기(120)가, 분석적 오디오 세그먼트(124)가 그에게 알려진 임의의 특정 사운드와 상관된다고 결정할 수 없는 경우, 그는 조정 메시지를 시간 수용 필터(114)에 전송할 수 있다. 특히, 조정 메시지는 시간 수용 필터(114)가 분석적 오디오 세그먼트들(124)의 각각 또는 전부에 대한 필터 시간을 조정하게 할 수 있다. 따라서, 분석적 오디오 세그먼트(124)가 너무 짧아서 특정 단음에 관한 충분한 의미 있는 정보를 캡처할 수 없는 경우, 시간 수용 필터는 분석적 오디오 세그먼트(124)의 길이 및/또는 바운드들(bounds)을 조정하여 단음을 더 잘 캡처할 수 있다. 따라서, 사운드 식별 단계를 갖지 않는 예시적인 실시예들에서도, 불확실성의 추정치들(estimates of uncertainty)이 생성되어 수용 필드를 조정하는 데 사용될 수 있다. 대안적으로, 한 번에 모두 동작하는 상이한 수용 필드들을 사용하는 다수의 기계 학습 시스템(116)(예를 들어, 음성 특징 추출기(120)의 서브컴포넌트들)이 있을 수 있고, 시스템의 나머지는 그들 각각으로부터의 결과들 사이에서 선택하거나 통합할 수 있다.

[0061] 특징 추출기(120)는 전체 수용 필드(114)에서 주파수 분포를 볼 필요가 없다. 예를 들어, 특징 추출기(120)는 제공된 수용 필드(114)보다 작은 수용 필드를 볼 수 있다. 또한, 시간 수용 필드(114)의 크기 및 폭(stride)은 기계 학습에 의해 조정될 수 있다.

[0062] 도 5b는 타겟(104)에 의해 발음된 단어 "Stella" 내의 "a" 단음에 대한 스펙트로그램(126)을 도시한다. 이 스펙트로그램(126)은 7개의 명확하게 가시적인 피크(128)를 갖는다. 물론, 주파수 데이터를 또한 갖는 다수의 다른 피크(128)가 있지만, 이들은 명확하게 가시적인 피크들(128)만큼 큰 강도를 갖지 않는다. 이러한 덜 가시적인 피크들은 타겟 음성(104)에 의해 이루어진 사운드 내의 고조파들(130)을 나타낸다. 이러한 고조파들(130)은 스펙트로그램(126)에서 인간에게 명확하게 인식 가능하지 않지만, 시스템(100)은 기본 데이터(underlying data)를 인지하고 그것을 사용하여 타겟 음성(104)에 대한 음성 프로파일을 생성하는 것을 돕는다.

[0063] 도 5c는 타겟(104)에 의해 발음된 단어 "Please" 내의 "ea" 단음에 대한 스펙트로그램(126)을 도시한다. 스펙트로그램(126)은 5개의 명확하게 가시적인 피크(128)를 갖는다. 도 5a 및 5b와 유사한 방식으로, 이 스펙트로

그램(126)은 또한 고조파 주파수들(130)을 갖는다. (예를 들어, 스펙트로그램들(126)에서) 주파수 데이터에 액세스함으로써, 시스템(100)은 특정 스펙트로그램(126)과 연관된 사운드를 결정한다. 또한, 이 프로세스는 스피치 샘플(105) 내의 다양한 분석적 오디오 세그먼트들(124)에 대해 반복된다.

[0064] 도 3으로 돌아가서, 프로세스는 단계 308로 진행하여, 타겟 음성(104)에 대한 벡터 공간(112)에서 부분 음성 프로파일을 매핑한다. 부분 음성 프로파일은 스피치 샘플(105) 내의 다양한 단음들의 주파수 분포들에 관한 데이터를 포함한다. 예를 들어, 도 5a-5c의 타겟(104)에 대해 도시된 3개의 단음에 기초하여 부분 음성 프로파일이 생성될 수 있다. 이 분야의 통상의 기술자는 이것이 부분 음성 프로파일의 실질적으로 단순화된 예인 것을 이해해야 한다. 일반적으로, 스피치 샘플(105)은 3개보다 많은 분석적 오디오 세그먼트(124)를 포함하지만, 더 적게 포함할 수 있다. 시스템(100)은 다양한 분석적 오디오 세그먼트들(124)에 대해 획득된 주파수 데이터를 취하고, 이들을 벡터 공간(112)에서 매핑한다.

[0065] 벡터 공간(112)은 연산들의 소정 세트가 양호하게 정의되는 데이터베이스 내의 벡터들이라고 하는 객체들의 집합을 지칭한다. 이러한 연산들은 해당 연산 하에서 결합(associativity), 교환(commutativity), 항등(identity) 및 역(inverse)과 같은 수학적 특성들을 따르는 벡터들의 가산; 및 해당 연산 하에서 호환(compatibility), 항등 및 분배(distributivity)의 수학적 특성들에 관한 스칼라들이라고 하는 객체들의 개별 클래스에 의한 승산을 포함한다. 벡터 공간(112) 내의 벡터는 통상적으로 N개의 숫자의 순서 리스트로서 표현되며, 여기서 N은 벡터 공간의 차원으로 알려져 있다. 이러한 표현이 사용될 때, 스칼라들은 통상적으로 단지 단일 숫자이다. 실수들의 3차원 벡터 공간에서, [1, -1, 3.7]은 예시적 벡터이고, 2*[1, -1, 3.7]=[2, -2, 7.4]는 스칼라에 의한 승산의 예이다.

[0066] 벡터 공간(112)의 예시적인 실시예들은 전술한 바와 같은 숫자들을 사용하지만, 통상적으로는 더 높은 차원의 사용예들이 있다. 특히, 예시적인 실시예들에서, 음색 벡터 공간(112)은 풍부 또는 선명과 같은 음색의 요소들을 나타내는 매핑을 지칭하며, 따라서 벡터들의 대응하는 요소들을 가산하거나 감산함으로써, 실제 음색의 소정 부분이 변경된다. 따라서, 타겟 음성(104)의 특성들은 벡터 공간에서 숫자들로 표현되며, 따라서 벡터 공간에서의 연산들은 타겟 음성(104)에 대한 연산들에 대응한다. 예를 들어, 예시적인 실시예들에서, 벡터 공간(112) 내의 벡터는 2개의 요소: [10 Hz 주파수의 진폭, 20 Hz 주파수의 진폭]를 포함할 수 있다. 실제로, 벡터들은 더 많은 수의 요소(예를 들어, 모든 가청 주파수 성분에 대한 벡터 내의 요소)를 포함할 수 있고/있거나, 더 세분화될 수 있다(finer-grained)(예를 들어, 1 Hz, 1.5 Hz, 2.0 Hz 등).

[0067] 예시적인 실시예들에서, 벡터 공간(112)에서 높은 피치 음성으로부터 낮은 피치 음성으로 이동하는 것은 모든 주파수 요소들을 수정하는 것을 필요로 할 것이다. 예를 들어, 이것은 여러 개의 높은 피치 음성을 함께 클러스터링하고, 여러 개의 낮은 피치 음성을 함께 클러스터링한 다음, 클러스터 중심들을 통과하는 라인에 의해 정의된 방향을 따라 이동함으로써 행해질 수 있다. 높은 피치 음성들의 몇 가지 예 및 낮은 피치 음성들의 몇 가지 예를 취하며, 이것은 공간(112)의 "피치" 액세스를 당신에게 제공한다. 각각의 음성은 다수의 차원(예를 들어, 32개의 차원)에 있을 수 있는 단일 벡터에 의해 표현될 수 있다. 하나의 차원은 기본 주파수의 피치일 수 있으며, 이는 대략적으로 남성 음성 및 여성 음성과 관련되고 이들을 구별한다.

[0068] 음성들(111)의 데이터베이스는 다양한 음성들에 대응하는 벡터 공간(112)에서 인코딩되는 벡터들을 보유한다. 이러한 벡터들은 벡터 공간(112)의 상황(context)에서 의미를 갖는 숫자들의 리스트들로서 인코딩될 수 있다. 예를 들어, 숫자들의 리스트의 제1 성분은 벡터 공간의 상황에서 "높은 피치 음성"을 의미할 수 있는 -2일 수 있거나, 벡터 공간의 상황에서 "낮은 피치 음성"을 의미할 수 있는 2일 수 있다. 기계 학습 시스템(116)의 파라미터들은 이러한 숫자들이 어떻게 처리될지를 결정하며, 따라서 생성기(140)는 리스트의 제1 성분에서 -2를 보는 것에 기초하여 입력 스피치를 높은 피치 음성으로 변환할 수 있거나, 음성 특징 추출기는 데이터베이스(111)에 저장된 숫자들의 리스트의 제2 성분에서 2를 갖는 벡터로서 낮은 피치 음성을 인코딩할 수 있다.

[0069] 예시적인 실시예들에서, 벡터 공간(112)은 통상적으로 전술한 특성들의 종류들을 나타낸다. 예를 들어, 깊은 음성 및 높은 피치 음성의 평균은 대략 중간 범위인 음성이어야 하고; 명확한 음성의 방향으로 약간 이동된 불명확한 음성(예를 들어, 명확한 음성에서 불명확한 음성을 감산하여, "불명확(gravelly)"에서 "명확(clear)"을 가리키는 벡터를 획득하고, 이것을 작은 스칼라와 승산하여 벡터가 조금만 변경되게 한 후에, 이것을 불명확한 음성에 더함)은 약간 더 명확하게 들려야 한다.

[0070] 스펙트로그램에 대해 수학적 연산들(예를 들어, 음성들의 평균화)을 수행하는 것은 자연스럽게 들리지 않는 사운드를 생성한다(예를 들어, 2개의 음성을 평균화하는 것은 2명의 사람이 한꺼번에 말하는 것처럼 들린다). 따라서, 스펙트로그램을 사용하여 깊은 음성 및 높은 피치 음성을 평균화하는 것은 중간 피치 음성을 산출하지 않

는다. 대조적으로, 벡터 공간(112)은 시스템(100)으로 하여금 중간 피치 음성을 생성하기 위해 높은 피치 음성과 낮은 피치 음성을 "평균화"하는 것과 같이 음성에 대해 수학적 연산들을 수행할 수 있게 한다.

[0071] 도 6a-6d는 본 발명의 예시적인 실시예들에 따른 벡터 공간(112)을 개략적으로 도시한다. 프로세스(300)는 결정(310)으로 진행하여, 이것이 벡터 공간에서 매핑된 제1 음성인지를 결정한다. 이것이 매핑된 제1 음성인 경우, 벡터 공간(112)에서의 그의 상대 위치는 중요하지 않다. 시스템(100)은 음성(104)을 비교할 상대 스케일(relative scale)이 없기 때문에 임의의 위치에서 벡터 공간(112)에서 음성(104)을 매핑할 수 있다.

[0072] 도 6a는 도 5b에 도시된 바와 같이 단어 "Stella"에서의 같은 "a" 사운드에 대한 타겟 음성(104)만을 포함하는 벡터 공간(112)을 개략적으로 도시한다. 예시적인 실시예들은 특정 사운드에 대한 벡터 공간(112)에 대한 도면들을 논의하고 도시하지만, 이 분야의 통상의 기술자는 벡터 공간(112)이 임의의 특정 사운드와 무관한 음색을 매핑한다는 것을 이해한다. 따라서, 특정 음성이 새로운 사운드를 말하는 것을 듣는 것은 벡터 공간(112)이 전체 벡터 공간(112)에 화자를 배치하는 것을 돕는다. 예시적인 실시예들은 기계 학습 시스템(116)이 음성들을 매핑할 수 있는 방식들을 간단히 예시하기 위해 특정 사운드들에 대한 벡터 공간들(112)을 도시하고 참조한다.

[0073] 타겟 음성(104)은 데이터베이스(112)에서 매핑된 제1(그리고 유일한) 음성이기 때문에, 데이터베이스(112) 전체는 타겟 음성(104)에만 관련된 정보를 반영한다. 따라서, 시스템(100)은 모든 음성들이 타겟 음성(104)이라고 간주한다. 이것이 제1 음성이기 때문에, 프로세스는 전술한 바와 같이 루프백(loop back)하고, 제2 음성을 매핑한다.

[0074] 도 7a는 제2(남성) 음성(132)에 의해 만들어진 단어 "Call" 내의 "a" 사운드에 대한 스펙트로그램(126)을 개략적으로 도시한다. 이것은 도 5a에서 타겟(104)에 의해 발음된 단음과 동일한 단음이다. 그러나, 제2 남성 음성(132)은 11개의 가시적인 피크(128)만을 갖는다. 또한, 제2 남성 음성(132)에 대한 가시적인 피크들(128)은 2 kHz 주파수를 초과하는 반면, 타겟(104)에 대한 가시적인 피크들(128)은 2 kHz 주파수보다 작다. (예를 들어, 스펙트로그램들(126)에 의해 표시된 바와 같은) 주파수 분포들의 차이에도 불구하고, 예시적인 실시예들에서, 상관 엔진(122)은 주파수 분포가 "call" 내의 단음 "a"를 나타내고, 이에 따라 벡터 공간(112)에서 그것을 매핑하는 것으로 결정할 수 있다. 예시적인 실시예들에서, 시스템(100)이 동일한 단음(예를 들어, 단어 "Call"에서와 같은 단음 "a")에 대한 다른 화자에 대한 데이터가 있는 것으로 결정한 후에, 시스템(100)은 예를 들어 이전에 설명된 프로세스들을 사용하여 벡터 공간(112)에서 서로에 대해 화자들을 매핑한다.

[0075] 도 6b는 타겟 음성(104)과 제2 남성 음성(132)을 매핑하는 단음: 단어 "Stella"에서와 같은 "a" 사운드에 대한 벡터 공간(112)을 개략적으로 도시한다. 시스템(100)은 타겟 음성(104) 및 제2 음성(132)에 의해 발음된 단음에 관한 데이터를 비교한다. 주파수 분포 특성들은 시스템(100)이 서로에 대해 음성들을 플로팅할 수 있게 한다. 따라서, 시스템(100)이 "a" 사운드의 완전히 새로운 입력을 수신하면, 시스템은 어느 음성이 입력에 대한 가장 유사한 주파수 특성들을 갖는지에 기초하여 타겟 음성(104)과 제2 음성(132) 사이를 구별할 수 있다.

[0076] 도 6b는 완전히 별개의 세그먼트들로서 매핑된 음성들(104, 132)을 도시하지만, 경계들은 그렇게 분명하지 않다는 것을 이해해야 한다. 사실상, 이러한 경계들은 특정 음성이 특정 주파수 분포를 나타낼 확률들을 나타낸다. 따라서, 실제로, 하나의 음성은 다른 음성의 도시된 구역에 중첩하는 사운드를 생성할 수 있다(예를 들어, 주파수 특성들의 중첩). 그러나, 음성 경계들은 특정 주파수 분포들을 갖는 사운드들이 특정 화자로부터 유래될 가장 큰 확률을 갖는다는 것을 보여주도록 의도된다.

[0077] 프로세스에서의 단계 310은 또한 매핑할 음성들이 더 있는지를 결정한다. 매핑할 음성들이 더 있다면, 단계 302-310이 반복된다. 도 7b는 제3 음성(여성)(134)에 의해 만들어진 단어 "Call" 내의 "a" 단음에 대한 스펙트로그램(126)을 개략적으로 도시한다. 이러한 제3 음성(134)은 6개의 가시적인 피크(128)를 갖는다. 제3 음성(134)에 대한 피크들(128)은 타겟 음성(104) 및 제2 음성(132)에서와 같이 압축되지 않는다. 다시, (예를 들어, 스펙트로그램들(126)에 의해 표시되는 바와 같은) 주파수 분포들의 차이에도 불구하고, 상관 엔진(122)은 주파수 분포가 "call" 내의 "a" 단음을 높은 확률로 나타내는 것으로 결정할 수 있다. 시스템(100)은 이 추가적인 음성을 벡터 공간(112)에서 매핑한다. 또한, 시스템(100)은 이제 도 6c에 도시된 바와 같이 3개의 화자 사이에서 단어 "call" 내의 "a" 사운드를 구별하기 위해 학습한다. 일부 실시예들에서, 음성 특징 추출기(120) 및 생성기(140)는 판별기(142)에 대해 적대적으로 역전파(backpropagation)를 통해 말단 대 말단 방식(end-to-end)으로 트레이닝된다.

[0078] 도 6d는 다양한 예들을 사용하는 프로세스(300)의 여러 사이클 후의 벡터 공간(112)을 도시한다. 복수의 음성이 벡터 공간(112)에서 매핑된 후에, 시스템(100)은 음성들을 더 정확하게 구별한다. 특정 화자에 기인하는 주

파수 특성들은 벡터 공간(112)이 비교할 더 많은 음색 데이터를 가짐에 따라 더 특유해진다. 음성들이 파선 원들로서 도시되지만, 원은 스펙트로그램들(126)에 도시된 바와 같은 주파수들의 복합 세트, 및 또한 그것의 변형들(이는 음색 "허용 오차"로서 설명될 수 있는데, 예를 들어 다양한 약간 변경된 주파수 분포들은 그것들이 동일한 음성으로부터 나오는 것처럼 들릴 수 있음)을 나타낸다는 것을 이해해야 한다.

[0079] 또한, 벡터 공간(112)은 소정 음색들과의 연관성을 형성하기 시작한다. 예를 들어, 특성 라인(136)이 생기기 시작하여 여성 음성들로부터 남성 음성들을 구별한다. 특성 라인(136)은 음성들을 완벽하게 구별하는 것으로 도시되지 않지만, 그것은 상당히 정확할 것으로 예상된다. 특정 음성의 음색 또는 주파수 분포들의 집합이 주로 생리학적 인자들에 의해 야기되기 때문에 특성들(characteristics)(예를 들어, 성별, 민족성, 나이 등)에 의해 음색들을 특성화하는 것이 가능하다. 특정 화자에 의해 만들어진 사운드들은, 그 형상이 사운드의 음색을 결정하는 후두 상위 성도(supralaryngeal vocal tract)에 의해 필터링된다. 성대들의 크기(예를 들어, 두께, 폭 및 길이)는 소정 진동들을 야기하며, 이는 상이한 주파수들, 따라서 상이한 음색들을 야기한다. 예를 들어, 여성은 유전적으로 남성보다 더 높은 포먼트 주파수들, 및 피크들(128) 사이의 더 큰 갭들을 갖는 성향이 있다. 따라서, 생리학적으로 유사한 모집단들(예를 들어, 남자 대 여자, 백인 대 아프리카계 미국인 등)은 특정 단음들에 관하여 더 유사한 주파수 분포들을 갖는다.

[0080] 단계 312에서, 프로세스는 또한 타겟 음성(104)에 대한 합성 음성 프로파일을 추정한다. 합성 음성 프로파일은 진정한 주파수 분포 데이터가 존재하지 않는 단음들에 대해 기계 학습 시스템(116)에 의해 예측된 주파수 분포들의 세트이다. 예를 들어, 도 5a-5c에 도시된 바와 같이, 시스템(100)은 문구 "CALL STELLA PLEASE" 내의 단음들에 관한 실제 데이터를 가질 수 있다. 그러나, 시스템(100)은 타겟 음성(104)으로부터의 Dog 내의 "D" 단음에 관한 어떠한 진정한 데이터도 갖지 않는다.

[0081] 도 8a는 본 발명의 예시적인 실시예들에 따른 합성 음성 프로파일(138)을 포함하는 벡터 공간(112)을 개략적으로 도시한다. 도시된 벡터 공간(112)은 "DOG" 내의 "D" 단음에 대한 것이다. 도 8은 "D" 단음을 만든 복수의 음성에 대한 매핑된 진정한 데이터를 도시한다. 도 6d에서 설명된 바와 같이, 타겟 음성(104)은 상이한 단음: "CALL"에서와 같은 "A"에 대한 이들 음성에 대해 매핑되었다. 다양한 단음들에 대한 주파수 분포들의 변화는 일반적으로 예측 가능하기 때문에, 기계 학습 시스템(116)은 어떠한 진정한 데이터도 존재하지 않는 단음들의 주파수 분포에 관한 예측들을 행한다. 예를 들어, 기계 학습은 "D" 단음에 대한 타겟 음성(104)의 합성 음성 프로파일(138)을 다른 음성들에 대해 매핑한다.

[0082] 합성 음성 프로파일(138)을 생성하기 위해, 타겟 음성(104)에 대한 부분 프로파일이 다른 저장된 음성 프로파일들과 비교되고, 타겟 음성(104)에 대한 합성 음성 프로파일(138)이 비교의 결과로서 추정된다. 따라서, 시스템(100)에 이전에 제공되지 않은 단음들은 타겟 음성(104)으로부터의 비교적 작은 스피치 샘플(105)로부터 추정될 수 있다. 예시적인 실시예들의 상세들이 이하에서 논의된다.

[0083] 초기 문제로서, 벡터 공간(112)은 복잡한 다차원 구조이고, 따라서 벡터 공간(112)의 2차원 슬라이스들은 도면들에서 특정 단음들에 대해 도시된다는 것을 이해해야 한다. 그러나, 도시된 다양한 단음 벡터 공간들(112)은 예시적인 목적을 위한 것일 뿐이며, 더 복잡한 3차원 벡터 공간(112)의 일부이다. 타겟 음성(104)에 대한 진정한 음성 프로파일에서의 주파수 분포들(예를 들어, 스피치 샘플(105)로부터의 모든 이용 가능한 단음 데이터에 대한 주파수 분포들)은 다른 매핑된 음성 프로파일들과 비교된다. 합성 음성 프로파일(138)은 누락된 단음들에 대해 추정된다. 이 분야의 통상의 기술자는 조정들이 특정 단음에 대한 음성 프로파일의 슬라이스에 대해 도시되지만, 실제로 조정은 도시하기 쉽지 않은 전체 다차원 음성 프로파일에 대해 이루어진다는 것을 이해할 것이다. 조정들은 신경망(116)과 같은 기계 학습 시스템(116)에 의해 달성될 수 있다.

[0084] 기계 학습 시스템(116)은 바람직하게는 자동화된 피드백 루프를 사용하여 그 자신을 최적화하고 당해 문제를 해결하는 자신의 능력을 개선하는 특수한 클래스의 문제 해결기(problem solver)이다. 기계 학습 시스템(116)은 그가 해결하려고 시도하는 실제 문제로부터 입력들을 취하지만, 완전히 그 자신에 내부적인 다양한 파라미터들 또는 설정들도 갖는다. 기계 학습 시스템(116)은 데이터 과학 시스템과 달리 다양한 입력들에 대한 그의 주어진 문제를 해결하려고 자동으로 시도하고, (항상은 아니지만, 때로는, 그의 답변들에 대한 자동화된 피드백의 도움으로) 미래의 시도들이 더 나은 결과들을 생성하도록 그의 파라미터들을 업데이트하도록 구성될 수 있다. 이 업데이트는 기계 학습 시스템(116)의 트레이닝 시작 이전에 선택되는 특정한, 수학적으로 잘 정의된 절차에 따라 발생한다.

[0085] 도면들을 참조하여 간단히 설명되었지만, 합성 음성(138)을 추정하는 것은 2개의 단음의 주파수 분포들을 비교하는 것만큼 간단하지는 않다. 타겟 음성(104)의 부분 음성 프로파일은 복수의 상이한 분석적 오디오 세그먼트

(124), 따라서 단음들에 관한 데이터를 포함한다. 상이한 단음들에 대한 주파수 분포의 변동들은 일반적인 경향들을 갖지만, 단음들 사이에는 범용 수학 공식/변환 비율(universal mathematical formula/conversion ratio)이 없다. 예를 들어, 단지 음성 A가 단음 "a"에 대해 음성 B 및 음성 C의 중간에 직접 속한다는 것은 음성 A가 단음 "d"에 대해 음성 B 및 음성 C의 중간에 직접 속한다는 것을 의미하지 않는다. 음성 분포들의 예측의 어려움은 이들이 복잡한 신호들(즉, 각각의 강도를 각각 갖는 주파수들의 범위)이라는 사실에 의해 심해진다. 또한, 특정 단음과 유사한 사운딩 음색을 제공할 수 있는 다수의 상이한 주파수 분포가 존재한다. 따라서, 기계 학습 시스템(116)은 특정 단음에 대한 주파수 분포들의 범위를 제공해야 하는 과제를 갖는다. 시스템(100)이 더 많은 음성을 매핑할수록, 일반적으로 합성 음성 프로파일(138)은 타겟 음성(104)의 음색과 더 양호하게 매칭된다.

[0086] 타겟 음성(104)을 벡터 공간(112)에 위치시키는 것을 돕기 위해, 생성기(140) 및 판별기(142)는 도 9를 참조하여 후술하는 피드백 루프를 실행할 수 있다. 일부 실시예들에서, 음성 특징 추출기(120)가 이전에 많은 음성에 대해 트레이닝된 경우(즉, 피드백 루프를 사용하여 이전에 많은 음성을 매핑한 경우), 타겟 음성(104)은 피드백 루프를 사용하지 않고 벡터 공간 내에 위치될 수 있다. 그러나, 다른 실시예들은 음성 특징 추출기(120)가 많은 음성에 대해 트레이닝되었더라도 피드백 루프를 여전히 사용할 수 있다.

[0087] 단계 314에서, 프로세스는 또한 합성 음성 프로파일(138)을 세밀화한다. 도 8b는 본 발명의 예시적인 실시예들에 따른, 합성 음성 프로파일(138)이 생성적 적대적 신경망(116)을 사용하여 세밀화된 후의 "DOG" 내의 단음 "D"에 대한 벡터 공간(112)을 개략적으로 도시한다. 생성적 적대적 신경망(116)은 생성적 신경망(140) 및 판별적 신경망(142)을 포함한다.

[0088] 생성적 신경망(140)은 그의 "문제"가 미리 정의된 클래스에 속하는 현실적인 예들을 생성하는 것인 일종의 기계 학습 시스템(116)이다. 예를 들어, 얼굴들에 사용되는 생성적 신경망은 현실적으로 보이는 얼굴들의 이미지들을 생성하려고 시도할 것이다. 예시적인 실시예들에서, 생성적 신경망(140)은 타겟 음색(104)의 스피치의 현실적인 예들을 생성한다.

[0089] 판별적 신경망(142)은 그의 "문제"가 그의 입력이 속하는 카테고리를 식별하는 것인 일종의 기계 학습 시스템(116)이다. 예를 들어, 판별적 신경망(142)은 이미지 설정들에서 개 또는 늑대의 사진들이 주어졌는지를 식별할 수 있다. 예시적인 실시예들에서, 판별적 신경망(142)은 입력된 스피치가 타겟(104)으로부터 유래된 것인지의 여부를 식별한다. 대안적으로 또는 추가로, 판별적 신경망(142)은 입력된 스피치의 화자를 식별한다.

[0090] 도 9는 본 발명의 예시적인 실시예들에 따른 증강된 음성 프로파일(144)을 세밀화하기 위해 생성적 적대적 망(116)을 사용하는 시스템(100)의 블록도를 도시한다. 증강된 음성 프로파일(144)은 기계 학습 시스템(116)에 의해 생성된 합성 음성 프로파일(138)에 더한, 스피치 샘플(105)로부터 획득된(진정한) 음성 프로파일의 조합이다. 벡터 공간(112)은 증강된 음성 프로파일(144)을 생성적 신경망(140)에 제공한다. 생성적 신경망(140)은 증강된 음성 프로파일(144)을 사용하여, 후보 스피치 세그먼트(146)(즉, 타겟(104)을 모방하는 것으로 되어 있지만, 타겟(104)으로부터의 진정한 스피치가 아닌 스피치)를 나타내는 스피치 데이터를 생성한다. 생성된 후보 스피치 세그먼트(146)는 후보 음성 내에 있다고 말할 수 있다. 후보 스피치 세그먼트(146)를 나타내는 스피치 데이터는 판별적 신경망(142)에 의해 평가되며, 이는 그가 믿는 바와 같이 후보 스피치 세그먼트(146) 내의 후보 음성을 나타내는 스피치 데이터가 진정한 또는 합성 스피치인지를 결정한다.

[0091] 시스템(100)이 오디오 후보 스피치 세그먼트(146)를 생성하면, 그것은 본질적으로 후보 스피치 세그먼트(146)를 나타내는 스피치 데이터를 포함한다. 그러나, 생성기(140)는 오디오 파일로서 결코 실제로 출력되지 않는 후보 스피치 세그먼트(146)를 나타내는 데이터를 제공할 수 있다. 따라서, 후보 스피치 세그먼트(146)를 나타내는 스피치 데이터는 파형, 스펙트로그램, 보코더 파라미터들, 또는 후보 스피치 세그먼트(146)의 운율학 및 단음 콘텐츠를 인코딩하는 다른 데이터로서의 오디오 형태일 수 있다. 또한, 스피치 데이터는 신경망(116)의 일부의 중간의 출력일 수 있다. 이러한 출력은 정상적인 인간 관찰자에 의해 이해되지 않을 수 있지만(예를 들어, 운율학 데이터 및 단음 데이터는 분리될 필요가 없음), 신경망(116)은 그러한 정보를 이해하고, 기계 학습(116) 또는 그의 부분들에 의해 이해가능한 방식으로 인코딩한다. 이하의 추가적인 논의는 편의상 "후보 스피치 세그먼트(146)"를 참조하지만, "후보 스피치 세그먼트(146)를 나타내는 더 넓은 스피치 데이터"를 포함하는 것으로 이해되어야 한다.

[0092] 예시적인 실시예들에서, 후보 스피치 세그먼트(146)는 소스 스피치 세그먼트(103)에 기초하여 생성된다. 도 1에서는 사용자(즉, 아놀드)인 것으로 도시되지만, 소스 음성(102)은 트레이닝 시에 시스템(100)에 입력될 필요가 없다. 소스 음성(102)은 시스템(100) 내에 입력되거나, 시스템(100)에 이미 저장되었거나, 시스템(100)에

의해 합성되는 임의의 음성일 수 있다. 따라서, 소스 스피치 세그먼트(103)는 사용자에 의해 제공될 수 있거나, 이미 시스템(100)에 있는 음성(예를 들어, 매핑된 음성)으로부터 스피치 세그먼트에 의해 제공될 수 있거나, 시스템(100)에 의해 생성될 수 있다. 사용자가 그들의 스피치를 변환하는 것, 생성된 음성 및/또는 시스템(100)에 이미 있는 스피치를 갖는 음성은 소스 음성(102)으로 간주될 수 있다. 또한, 도 9에 도시된 피드백 루프 동안 상이한 후보 스피치 세그먼트들(146)이 생성됨에 따라, 상이한 소스 스피치 세그먼트들(103)이 사용될 수 있다.

[0093] 판별적 신경망(142)은 후보 스피치 세그먼트(146), 및 또한 타겟 음성(104)을 포함하는 복수의 음성에 관한 데이터를 수신한다. 예시적인 실시예들에서, 생성기(140) 및 판별기(142)는 타겟 음성을 포함하는 복수의 음성 프로파일에 관한 데이터를 수신한다. 이것은 신경망(116)이 다른 음성들의 복수의 음색 데이터를 참조하여 스피치가 다소 타겟(104)과 같이 들리게 하는 변경들을 식별할 수 있게 한다. 그러나, 타겟 음성(104) 자체에 관한 데이터는 복수의 음성과 암시적으로 관련될 수 있다는 것이 이해되어야 하는데, 이는 다른 음성들의 특성들이 타겟 음성(104)을 매핑하거나 세밀화할 때 판별기(142)의 학습된 파라미터들을 통해 이미 어느 정도 이해되었기 때문이다. 또한, 타겟 음성(104)이 트레이닝 또는 벡터 공간(112)에 대한 더 많은 음성의 추가를 통해 세밀화됨에 따라, 타겟 음성(104)은 복수의 음성에 대한 데이터를 더 제공한다. 따라서, 예시적인 실시예들은 생성기(140) 및/또는 판별기(142)가 복수의 음성 프로파일로부터 데이터를 명시적으로 수신하는 것을 요구할 수 있지만, 요구하지 않는다. 대신에, 생성기(140) 및/또는 판별기(142)는 복수의 음성 프로파일에 기초하여 수정된 타겟 음성(104) 프로파일로부터의 데이터를 수신할 수 있다. 이전 시나리오들 중 어느 하나에서, 시스템(100)은 복수의 음성 프로파일을 참조하여 데이터를 수신한다고 말할 수 있다.

[0094] 예시적인 실시예들에서, 생성기(140)는 타겟(104) 이외의 음성처럼 들리는 후보 스피치 세그먼트들(146)의 생성에 (판별기(142)에 의해) 제재가 가해진다. 예시적인 실시예들에서, 생성기(140), 음성 특징 추출기(120) 및/또는 판별기(142)는 복수의 음성 프로파일에 관한 데이터에 액세스한다. 따라서, 생성기(140), 판별기(142) 및/또는 음성 특징 추출기(120)는 복수의 상이한 음성의 음색 데이터를 참조하여 결정을 행할 수 있다. 따라서, 생성기(140)는, 해당 화자가 타겟(104)과 매우 유사하더라도, 타겟(104) 이외의 누군가와 같이 합성 스피치가 들리게 하는 타겟 음성(104) 프로파일에 대한 변경을 행하지 않는다. 생성기(140)는 복수의 음성 프로파일에 관한 데이터에 대한 액세스를 갖기 때문에, 그것은 타겟과 잠재적으로 유사하게 들리는 다른 화자들을 구별할 수 있어서, 더 나은 품질의 후보 스피치 세그먼트들(146)을 생성할 수 있다. 이어서, 판별기(142)는 더 세밀한 상세들을 픽업하고 더 상세한 불일치 메시지(148)를 제공한다. 도면에 도시되지 않았지만, 불일치 메시지(148)는 음성 특징 추출기(120)에 제공될 수 있고, 이는 이어서 벡터 공간(112)에서 음성 프로파일들을 수정한다.

[0095] 전술한 바와 같이, 판별적 신경망(142)("판별기(142)"라고도 함)은 후보 스피치 세그먼트(146)가 타겟(104)으로부터 유래된 것인지의 여부를 식별하려고 시도한다. 이 분야의 통상의 기술자는 후보 스피치 세그먼트(146)가 타겟 음성(104)으로부터 유래된 것인지를 결정하기 위해 사용될 수 있는 상이한 방법들을 이해한다. 특히, 판별기(142)는 소정의 주파수들 및/또는 주파수 분포들이 타겟 음성(104)의 음색의 일부일 가능성이 있는지 여부를 결정한다. 판별기(142)는 후보 스피치 세그먼트(146)를 벡터 공간(112)에서 매핑된 타겟 음색(104) 및 다른 음성들과 (즉, 복수의 상이한 음성의 복수의 음색 데이터를 참조하여) 비교함으로써 이것을 행할 수 있다. 따라서, 벡터 공간(112)에서 더 많은 음성이 매핑될수록, 판별기(142)는 더 양호하게 합성 스피치로부터 진정한 스피치를 식별한다. 따라서, 일부 실시예들에서, 판별기(142)는 아이덴티티를 후보 음성 및/또는 후보 스피치 세그먼트(146)에 할당할 수 있다.

[0096] 예시적인 실시예들에서, 판별기(142)는 케이던스, 액센트 등과 같은 것들에 기초하여 그가 "보고"/판별하는 것을 방지하는 시간 수용 필드(114)를 갖는다. 추가로 또는 대안적으로, 생성기(140)는 케이던스, 액센트 등과 같은 것들에 기초하여 그가 생성하는 것을 방지하는 시간 수용 필드(114)를 갖는다. 따라서, 후보 스피치 세그먼트(146)는 케이던스, 액센트 등과 같은 더 긴 시간 특성들을 포함하는 것을 회피하기에 충분히 짧게 생성될 수 있고/있거나, 시간 수용 필드(114)를 사용하여 필터링될 수 있다. 따라서, 판별기(142)는 이러한 다른 특성들에 기초하여 판별하기보다는, 음색에 기초하여 가짜 스피치로부터 진정한 스피치를 구별한다.

[0097] 판별기(142)는, 예를 들어, 소정의 단음들의 기본 주파수를 비교하여 어느 가능한 음색이 가장 명확하게 매치인지(즉, 매치일 가장 높은 확률을 갖는지)를 파악함으로써 시작할 수 있다. 전술한 바와 같이, 기본 주파수 이외에 음색을 정의하는 더 많은 특성이 있다. 시간이 지남에 따라, 판별기(142)는 음성을 식별하는 더 복잡한 방식들을 학습한다.

[0098] 본 발명자들에게 알려진 종래 기술의 스피치 대 스피치 변환 시스템들은 열악한 품질의 변환들을 생성한다(예를

들어, 오디오는 타겟 음성처럼 들리지 않는다). 대조적으로, 예시적인 실시예들은 생성적 신경망(140)("생성기(140)"로도 지칭됨) 및 판별기(142)가 타겟 음성(104)만이 아니라 그 이상의 것을 사용하여 트레이닝되기 때문에 상당히 더 높은 품질의 변환들을 생성한다. 예를 들어, 종래 기술의 시스템은 일본 여성으로부터의 스피치를 버락 오바마의 음성으로 변환하려고 시도될 수 있다. 그러한 종래 기술의 시스템은 그가 할 수 있는 만큼 버락 오바마에 가까워지지만, 다른 음성들과 비교하는 방법에 관계없이 그렇게 한다. 이러한 종래 기술의 시스템은 사람들이 상이한 사람 음성들을 구별하는 방법을 이해하지 못하기 때문에, 종래 기술의 생성기는 종래 기술의 판별기를 속이기 위해 실제로 음성이 그의 탐색 내의 다른 누군가의 음성에 더 가깝게 들리게 하는 트레이드오프를 행할 수 있다.

[0099] 판별기(142)가 차이를 검출하지 못하는 경우, 프로세스는 종료된다. 그러나, 판별기(142)가 후보 스피치 세그먼트(146)가 타겟 음성(104)으로부터 유래되지 않은 것을 검출하는 경우(예를 들어, 후보 음성이 타겟 음성과 다른 경우), 불일치 메시지(148)가 생성된다. 불일치 메시지(148)는 판별기(142)가 후보 스피치 세그먼트(146)가 타겟 음색(104) 내에 있지 않은 것으로 결정한 이유에 관한 상세를 제공한다. 판별기(142)는 후보 스피치 세그먼트(146)를 (타겟(104)을 포함하는) 복수의 음성과 비교하여, 후보 스피치 세그먼트(146)가 타겟 음성(104) 내에 있는지를 결정한다. 예를 들어, 벡터 공간(112)에서 매핑된 복수의 음성에 의해 정의되는 인간 스피치의 소정 파라미터들을 비교함으로써, 불일치 메시지(148)는 후보 스피치 세그먼트(146)가 인간 스피치의 정확한 파라미터들 내에 있는지, 또는 정상적인 인간 스피치인 것의 밖에 속하는지를 결정할 수 있다. 또한, 벡터 공간(112)에서 매핑된 복수의 음성을 비교함으로써, 불일치 메시지(148)는 타겟 음성(104) 이외의 음성으로부터 유래될 더 높은 확률을 갖는 특히 주파수 데이터에 대한 상세를 제공할 수 있다. 따라서, 벡터 공간(112)은 이러한 불일치 메시지(148)를 피드백으로서 사용하여 타겟(104)의 증강된 음성 프로파일(144) 및/또는 합성 음성 프로파일(138)의 부분들을 조정할 수 있다.

[0100] 불일치 메시지(148)는, 예를 들어, 피크들(128)의 수, 특정 피크들(128)의 강도, (도 5a에서의) 공격(129), (도 5c에서의) 감쇠(131), 고조파들(130), 기본 주파수, 포먼트 주파수, 및/또는 시스템(100)이 타겟 음색(104)으로부터 후보 스피치 세그먼트(146)를 구별하는 것을 가능하게 하는 단음들 및/또는 분석적 오디오 세그먼트들(124)의 다른 특성들에서의 불일치들(예를 들어, 타겟 음성(104)으로부터 유래되지 않을 높은 확률을 갖는 주파수 데이터)에 관한 정보를 제공할 수 있다. 불일치 메시지(148)는 매우 복잡한 조합들에서 파형들의 임의의 특징에 효과적으로 대응할 수 있다. 불일치 메시지(148)는, 예를 들어, 네 번째 가장 큰 진폭 주파수가 "수상한" 진폭을 갖는 것으로, 그리고 그것이 진정한 것으로 보이게 하기 위해 그것으로부터 소정 양이 차감되어야 하는 것으로 결정할 수 있다. 이것은 불일치 메시지(148)에서 이용 가능한 정보의 종류를 예시하기 위한 매우 단순화된 예이다.

[0101] 벡터 공간(112)은 불일치 메시지를 수신하고, 그것을 사용하여 합성 음성 프로파일(138)(및 그 결과, 증강된 음성 프로파일(144))을 세밀화한다. 따라서, 도 8b에 도시된 바와 같이, 벡터 공간(112)은 타겟 음성 음색(104)에 할당되는 주파수 분포들의 세트를 좁히고/좁히거나 조정한다. 불일치 메시지(148)는 복수의 음색 데이터를 참조하여 후보 스피치 세그먼트(146)와 타겟 음색(104) 간의 불일치를 결정한다. 예를 들어, 타겟 음성(104)은 더 이상 코너 맥그리거 또는 버락 오바마와 중첩되지 않는다. 이 분야의 통상의 기술자는 신경망(116)이 음성들 간의 명확한 구별을 넘어서 계속 개선될 수 있다(예를 들어, 벡터 공간(112) 내의 대표적인 원을 좁힐 수 있다)는 것을 이해해야 한다. 판별기(142)는 화자를 식별하지만, 또한 (스피치가 생성기(140)에 의해 합성적으로 생성되더라도) 후보 스피치 세그먼트(146)가 진정한 스피치일 높은 확률을 갖는지를 결정하기 위해 단계를 더 진행한다. 예를 들어, 주파수 특성들이 특정 타겟에 가까운 경우에도(예를 들어, 화자 A의 확률이 90 퍼센트이고, 화자 B의 확률이 8 퍼센트이고, 나머지 화자들 사이에 분산된 확률이 2 퍼센트인 경우에도), 판별기(142)는 주파수 특성들이 임의의 인식 가능한 인간 스피치를 생성하지 않고 합성이라고 결정할 수 있다. 벡터 공간(112)은 이 데이터를 사용하여, 증강된 음성 프로파일(144)의 바운드들을 더 잘 정의하는 것을 돕는다.

[0102] 복수의 음성을 참조하여 증강된 음성 프로파일(144)을 세밀화하는 것은 종래 기술의 방법들에 비해 개선들을 제공한다. 이러한 개선들은 개선된 음성 변환 품질을 포함하며, 이는 사용자들이 공지된 종래 기술의 방법들을 사용하여 이용 가능하지 않은 현실적 음성 변환들을 생성할 수 있게 한다. 단일 음성(예를 들어, 타겟 음성)만 가지고 생성적 적대적 망(116)을 사용하는 것은 개선된 피드백(예를 들어, 불일치 메시지(148))을 유발하는 현실적인 문제 세트들(후보 스피치 세그먼트(146))을 생성하기에 충분한 데이터를 생성적 적대적 신경망(116)에 제공하지 못한다. 개선된 피드백은 시스템(100)이 궁극적으로 훨씬 더 사실적인 음성 변환을 제공할 수 있게 한다. 일부 실시예들에서, 판별기(142)가 후보 음색과 타겟 음색 사이의 임의의 차이들을 검출하지 못하는 경우, 차이가 결정되지 않았음을 나타내는 널 불일치 메시지(null inconsistency message)가 생성될 수 있다. 널

불일치 메시지는 피드백 프로세스가 종료될 수 있다는 것을 나타낸다. 대안적으로, 시스템(100)은 단순히 불일치 메시지를 생성하지 않을 수 있다.

[0103] 수정된 증강된 음성 프로파일(144)은 다시 생성적 신경망(140)으로 전송되고, 판별기(142)에 의한 고려를 위해 다른(예를 들어, 제2) 후보 스피치 세그먼트(146)가 생성된다. 제2 후보 스피치 세그먼트(146)(기타 등등)는 제2 후보 음성(기타 등등) 내에 있다고 말할 수 있다. 그러나, 일부 실시예들에서, 제1 후보 음성 및 제2 후보 음성은 반복마다 매우 유사한 사운딩일 수 있다. 일부 실시예들에서, 판별기(142)는 불일치 메시지(148)가 사소한 차이들을 검출할 수 있도록 정밀하게 조정될 수 있다. 따라서, 제1 후보 음성 및 제2 후보 음성은 인간 관찰자와 매우 유사하게 들릴 수 있지만, 여전히 이 논의의 목적을 위해 상이한 음성들로 간주될 수 있다.

[0104] 프로세스는 판별기가 후보 스피치 세그먼트(146)를 타겟 음색(104)으로부터 구별할 수 없을 때까지 계속된다. 따라서, 시간이 지남에 따라, 증강된 음성 프로파일(144)과 타겟 음성(104)의 실제 스피치 간의 차이는 판별기(142)에 의해 식별 가능하지 않아야 한다(예를 들어, 후보 스피치 세그먼트(146)가 타겟 음성(104)으로부터 유래될 확률은, 소정 실시예들에서는 더 낮은 퍼센트들도 충분할 수 있지만, 99+ 퍼센트로 개선될 수 있다). 타겟 음성(104)의 증강된 음성 프로파일(144)이 충분히 세밀화된 후에, 사용자들은 그들의 스피치 세그먼트(103)를 타겟 음성(104)으로 변환할 수 있다.

[0105] 도 8c는 제2 음성(132) 및 제4 음성의 추가를 갖는 도 8b의 벡터 공간(112)을 개략적으로 도시한다. 벡터 공간(112)에 대한 더 많은 음성의 추가는 음성들을 구별하기 위한 판별기(142)의 능력을 더 향상시킬 수 있다는 점에 유의해야 한다. 예시적인 실시예들에서, 제2 음성(132) 및 제4 음성으로부터의 데이터는 타겟 음성(104)에 대한 합성 음성 프로파일(138)을 세밀화하기 위해 사용된다. 또한, 제2 음성(132) 및 제4 음성은 코너 맥그리거와 같은 다른 화자들의 주파수 분포들을 세밀화하는 것을 도울 수 있다.

[0106] 도 3으로 돌아가서, 프로세스(300)는 매핑할 음성이 더 존재하는지를 결정하는 단계 316에서 종료된다. 존재하는 경우, 필요한 것만큼 여러 번 전체 프로세스가 반복된다. 합성 음성 프로파일(138)(즉, 가능한 주파수 분포들 및 따라서 음성의 사운드)은 일반적으로 벡터 공간(112)으로의 더 많은 음성의 추가에 의해 개선된다. 그러나, 매핑할 다른 음성이 없다면, 프로세스는 완료된다.

[0107] 예시적인 실시예들은 이전에 듣지 못한 완전히 새로운 음성들 및 음성들의 다양한 조합들을 생성한다. 특성 라인(136)을 참조하여 설명된 바와 같이, 기계 학습 시스템(116)은 벡터 공간(112)에서 매핑된 음성들에 대한 소정의 조직적인 패턴들을 생성하기 시작한다. 예를 들어, 유사한 성별, 인종 및/또는 나이의 음성들은 유사한 주파수 특성들을 가질 수 있고, 따라서 함께 그룹화된다.

[0108] 전술한 바와 같이, 벡터 공간(112)은 그 안의 데이터 세트들에 대한 수학적 연산들을 허용한다. 따라서, 예시적인 실시예들은 알 파치노 및 제임스 얼 존스의 음성들 사이의 음성과 같은 벡터 공간(112)에서의 수학적 연산들을 제공한다. 또한, 음성 생성 엔진은 또한 새로운 음성들을 생성하기 위해 그룹핑들(groupings)에 관한 일반화를 사용할 수 있다. 예를 들어, 새로운 음성은 평균 중국 여성 음성으로부터 평균 여성 음성을 빼고, 평균 남성 음성을 더함으로써 생성될 수 있다.

[0109] 도 10은 본 발명의 예시적인 실시예들에 따른 스피치 대 스피치 변환을 위한 프로세스(1000)를 도시한다. 이 프로세스는 스피치 대 스피치 변환을 위해 통상적으로 사용되는 더 긴 프로세스로부터 실질적으로 단순화된다는 점에 유의해야 한다. 따라서, 스피치 대 스피치 변환 프로세스는 이 분야의 통상의 기술자들이 사용할 가능성이 있는 많은 단계를 갖는다. 또한, 단계들 중 일부는 도시된 것과 상이한 순서로 또는 동시에 수행될 수 있다. 따라서, 이 분야의 통상의 기술자들은 프로세스를 적절하게 수정할 수 있다.

[0110] 프로세스는 단계 1002에서 시작하여, 스피치 세그먼트(103)를 나타내는 스피치 데이터를 시스템(100)에 제공한다. 예를 들어, 스피치 세그먼트(103)를 나타내는 스피치 데이터를 본질적으로 포함하는 스피치 세그먼트(103)가 입력(108)에 제공될 수 있다. 대안적으로, 생성기(140)는 (예를 들어, 텍스트 입력으로부터) 스피치 세그먼트를 나타내는 데이터를 제공할 수 있다. 따라서, 스피치 세그먼트(103)를 나타내는 스피치 데이터는 파형, 스펙트로그램, 보코더 파라미터들, 또는 스피치 세그먼트(103)의 운율학 및 단음 콘텐츠를 인코딩하는 다른 데이터로서의 오디오의 형태일 수 있다. 또한, 스피치 데이터는 신경망(116)의 일부의 중간의 출력일 수 있다. 이러한 출력은 정상적인 인간 관찰자에 의해 이해되지 못할 수 있지만(예를 들어, 운율학 데이터 및 단음 데이터는 분리될 필요가 없음), 신경망(116)은 그러한 정보를 이해하고, 기계 학습(116) 또는 그의 부분들에 의해 이해 가능한 방식으로 그것을 인코딩한다. 전술한 바와 같이, 스피치 세그먼트(103)는 사람의 스피치로부터 올 필요가 없고, 그 대신에 합성될 수 있다. 아래의 추가적인 논의는 편의상 "스피치 세그먼트(103)"를 참조하지

만, "스피치 세그먼트(103)를 나타내는 더 넓은 스피치 데이터"를 포함하는 것으로 이해되어야 한다.

[0111] 단계 1004에서, 사용자는 타겟 음성(104)을 선택한다. 타겟 음성(104)은 도 3을 참조하여 설명된 프로세스를 사용하여 벡터 공간(112)에서 이전에 매핑되었을 수 있다. 대안적으로, 새로운 음성이 도 3을 참조하여 설명된 프로세스를 또한 사용하여 시스템에 매핑될 수 있다. 스피치 세그먼트(103)가 입력되는 예시적인 실시예들에서, 스피치 세그먼트(103)는 타겟 음성(104)을 매핑하는 것을 돕기 위해 사용될 수 있지만, 그럴 필요는 없다(예를 들어, 후보 스피치(146)는 스피치 세그먼트(103)의 단음들, 액센트 및/또는 케이던스를 반영할 수 있다). 단계 306에서, 타겟(104)에 대한 증강된 음성 프로파일(144)이 취해지고 스피치 세그먼트(103)에 적용된다. 즉, 스피치 세그먼트(103)의 주파수는 타겟 음성(104)에 존재하는 주파수 분포들을 반영하도록 변환된다. 이것은 스피치 세그먼트(103)를 타겟 음성(104)으로 변환한다.

[0112] 적대적 트레이닝 동안, 생성적 신경망(140)은 (도 1의 런타임에 행하는 것처럼) 입력 스피치를 취하고 타겟 음색을 적용하지만, 판별기(142)는 출력 스피치를 보고, 타겟 음성(104)에 의해 (정의상, 판별기가 그것이 실제라고 믿더라도, 스피치는 합성일 것이지만) 그것이 "실제" 인간 스피치인지에 대한 결정을 행한다는 점에 유의해야 한다. 이와 달리, 도 1에 도시된 음성 변환 동안, 시스템(100)은 (추가 트레이닝이 선택적이지만) 추가 트레이닝에 대한 필요 없이 변환이 다소 원활하게 발생하여, 실시간 또는 거의 실시간 스피치 대 스피치 변환들이 발생할 수 있도록 충분한 음성들에 대해 이미 트레이닝되었다. 해당 타겟 화자에 의한 실제 인간 스피치의 트레이닝 세트 예들은 (입력 화자와 같은) 임의의 다른 화자에 의한 임의의 "오염"을 갖지 않으며, 따라서 생성적 신경망(140)은 입력 화자의 음색을 제거하고 타겟 화자의 음색을 대신 사용하도록 학습하며, 그렇지 않을 경우, 판별기(142)는 속지 않는다.

[0113] 단계 308에서, 변환된 스피치 세그먼트(106)는 타겟 음성(104)에서 출력된다. 그 다음, 단계 310에서 프로세스는 변환될 스피치 세그먼트들(103)이 더 존재하는지를 묻는다. 스피치 세그먼트들(103)이 더 존재하는 경우, 프로세스(1000)가 반복된다. 그렇지 않으면, 프로세스가 완료된다.

[0114] 일부 실시예들에서, 타겟(104) 화자는 사전-스크립트된 스피치 샘플(105)을 제공하도록 요청받을 수 있다. 예를 들어, 전부는 아니더라도 많은 일반적으로 발음된 단음들을 캡처하는, 타겟이 읽도록 요청되는 스크립트가 존재할 수 있다. 따라서, 예시적인 실시예들은 모든 단음에 대한 진정한 주파수 분포 데이터를 가질 수 있다. 또한, 예시적인 실시예들에서, 벡터 공간(112)은 적어도 하나, 바람직하게는 더 많은 음성으로부터의 모든 단음에 대한 진정한 주파수 분포 데이터를 갖는다. 따라서, 예시적인 실시예들은 진정한 데이터에 적어도 부분적으로 기초하여 합성 음성 프로파일들(138)을 추정할 수 있다.

[0115] 예시적인 실시예들은 스피치 샘플(105)을 타겟 "음성"(104) 내에 있는 것으로 참조하지만, 예시적인 실시예들은 발음된 단어들 및/또는 인간 음성들로 제한되지 않는다는 것을 이해해야 한다. 예시적인 실시예들은 악기, 로봇 및/또는 동물들에 의해 생성되는 것들과 같은 스피치 샘플(105) 내의 단음(인간의 단어 자체의 일부가 아님)만을 필요로 한다. 따라서, 예시적인 실시예들에서, 스피치 샘플(105)은 오디오 샘플(105)로도 지칭될 수 있다. 이러한 사운드들은 시스템에 의해 분석되고, "사운드 프로파일"을 생성하도록 매핑될 수 있다.

[0116] 예시적인 실시예들은 종래 기술에 비해 다수의 장점을 제공한다는 것을 더 이해해야 한다. 실시간 또는 거의 실시간 음성 변환은 타겟 음성(104)의 비교적 작은 스피치 샘플(105)로부터 가능해진다. 음성 대 음성 변환은 엔터테인먼트, (예를 들어, 가청 애플리케이션에서의) 오디오북 음성들의 변환, 개인용 음성 어시스턴트들(예를 들어, 아마존의 알렉사)의 맞춤화, 영화들을 위한 죽은 배우들(예를 들어, 스타워즈의 프린세스 리아)의 음성들의 재생 또는 (예를 들어, 고유한 음성 또는 죽은 가족 구성원의 음성을 갖기 위한) 인공 지능 로봇들을 위해 유용할 수 있다. 다른 용도들은 사용자들이 그들의 스피치의 부분들을 수정할 수 있는 "음성용 포토샵", 또는 상이한 노래/기구 부분들을 생성하고 이들을 단일 밴드/음성에 함께 넣기 위해 임의의 사운드 입력을 사용하는 "자동 밴드"를 포함할 수 있다. 다른 용도들은 동물들이 "대화(talk)"하게 하는 것, 즉 사람의 스피치를 특정 동물의 음색으로 변환하는 것을 포함한다.

[0117] 도 11은 본 발명의 예시적인 실시예들에 따른 음성을 사용하여 아이덴티티를 검증하는 프로세스를 도시한다. 위에서 논의된 다른 프로세스들과 같이, 이 프로세스는 음성을 사용하여 아이덴티티를 검증하기 위해 일반적으로 사용되는 더 긴 프로세스로부터 실질적으로 단순화된다는 점에 유의해야 한다. 따라서, 음성을 사용하여 아이덴티티를 검증하는 프로세스는 이 분야의 통상의 기술자들이 사용할 가능성이 있는 많은 단계를 갖는다. 또한, 단계들 중 일부는 도시된 것과 다른 순서로 또는 동시에 수행될 수 있다. 따라서, 이 분야의 통상의 기술자들은 프로세스를 적절하게 수정할 수 있다.

[0118] 프로세스(1100)는 단계 1102에서 시작하여, 복수의 매핑된 음성을 갖는 벡터 공간(112)을 제공한다. 벡터 공간은 전술한 바와 같이 복수의 음성으로 채워질 수 있다. 바람직하게는, 벡터 공간(112)은 1000개보다 많은 음성으로 채워지고, 음성들 각각은 50개 이상의 단음에 대해 매핑되었다.

[0119] 단계 1104에서, 방법은 아이덴티티가 검증되고 있는 사람으로부터 입력 스피치를 수신한다. 기계 학습 시스템(116)이 후보 스피치(146)가 타겟(104)에 대해 진짜인지를 결정하는 방법과 유사한 방식으로, 기계 학습 시스템(116)은 또한 임의의 입력 스피치가 아이덴티티가 검증되고 있는 사람에 대해 진짜인지를 결정할 수 있다. 단계 1106에서, 아이덴티티가 검증되고 있는 사람에 대해 진정한 음성 프로파일이 생성된다. 전술한 바와 같이, 음성 프로파일은 시간 수용 필드(114)를 사용하여 분석적 오디오 세그먼트들(124)을 필터링함으로써 생성될 수 있다. 변환 엔진(118)은 분석적 오디오 세그먼트들(124)의 주파수 성분들을 추출할 수 있고, 주파수 대 사운드 상관 엔진(122)은 특정 분석적 오디오 세그먼트 내의 주파수 성분들을 특정 사운드와 상관시킬 수 있다. 기계 학습(116)은 이어서 타겟 음성(104)의 진정한 음성 프로파일을 데이터베이스(112)에서 매핑할 수 있다.

[0120] 단계 1108에서, 프로세스(1100)는 진정한 음성 프로파일(및/또는 생성된 경우에는 증강된 음성 프로파일(144))을 벡터 공간(112) 내의 음성 프로파일들과 비교한다. 유사하게, 벡터 공간(112)에서 매핑되는 임의의 음성은 또한 진정한 음성 프로파일 및/또는 증강된 음성 프로파일(144)에 기초하여 검증될 수 있다. 비교에 기초하여, 기계 학습 시스템(116)은 존재할 경우에 벡터 공간(112) 내의 어느 음성이 해당 아이덴티티의 음성에 대응하는지를 결정할 수 있다. 따라서, 단계 1110에서, 프로세스는 해당 아이덴티티의 아이덴티티를 검증 및/또는 확인한다.

[0121] 단계 1112는 아이덴티티가 검증되는지를 묻는다. 예시적인 실시예들에서, 음성이 주파수 분포에 기초하여 95 퍼센트(예를 들어, 판별기가 95 퍼센트 신뢰 구간을 제공하는 경우) 이상의 매치인 경우, 음성이 검증된다. 일부 실시예들에서, 음성은 검증되기 위해 ("매치"라고 하는) 시스템 내의 다른 음성들과 비교하여 스피치가 아이덴티티 음성에 대응하는 적어도 99 퍼센트 신뢰도를 가져야 할 수 있다. 일부 다른 실시예들에서, 음성은 검증되기 위해 적어도 99.9 퍼센트 매치를 가져야 할 수 있다. 추가의 실시예들에서, 음성은 검증되기 위해 적어도 99.99 퍼센트의 매치를 가져야 할 수 있다. 음성이 검증되지 않으면, 프로세스는 음성의 다른 샘플을 수신할 것을 요청할 수 있고, 단계 1104로 복귀한다. 그러나, 음성이 검증되는 경우, 프로세스(1100)는 단계 1114로 진행하여, 액션을 트리거한다.

[0122] 단계 1114에서 트리거된 액션은 예를 들어 패스워드를 잠금 해제하는 것일 수 있다. 시스템(100)은 음성들을 비교하고 특정 스피치의 진정성/아이덴티티를 결정할 수 있다. 따라서, 시스템(100)은 음성 패스워드들의 사용을 가능하게 한다. 예를 들어, 아이폰 모바일 전화의 더 새로운 버전은 전화를 잠금 해제하기 위해 (예를 들어, 얼굴 인식 및/또는 지문 스캐닝에 더하여 또는 대안으로) 음성 검증을 이용할 수 있다. 시스템(100)은 음성을 분석하고(예를 들어, 그것을 벡터 공간(112)에서 애플에 의해 이전에 매핑된 다수의 음성과 비교하고), 음성이 매치인 경우 스마트폰을 잠금 해제한다. 이것은 사용 및 보안의 용이함을 증가시킨다.

[0123] 예시적인 실시예들에서, 트리거된 액션은 음성이 스마트 홈 애플리케이션들의 제어를 위한 허가를 갖는다는 신호를 잠금 해제 및/또는 제공한다. 예를 들어, 문을 잠금 및/또는 잠금 해제하고, 주방 기구들을 턴온하고, 기타 등등을 행하기 위한 커맨드들이 모두 적절한 액세스를 갖는 음성(예를 들어, 소유자)으로부터의 것인 것으로 검증되고 확인될 수 있다. 예시적인 실시예들은 스마트 홈 어시스턴트(예를 들어, 아마존 알렉사)에 통합될 수 있고, 커맨드들의 검증을 허용할 수 있다. 이것은 사용자의 음성을 확인함으로써 은행 이체, 대형 이체, 또는 개인 정보(예를 들어, 의료 기록)에 대한 액세스와 같은 민감한 기술들에 대한 아마존 알렉사의 사용을 가능하게 하는 것을 포함한다.

[0124] 또한, 예시적인 실시예들은 아이덴티티의 쉬운 검증을 위해 식별 시스템들(예를 들어, 경찰 및/또는 공항) 및 포스(point of sale) 시스템들(예를 들어, 상점의 등록기)에 통합될 수 있다. 따라서, 포스 시스템들에서, 트리거된 액션은 사용자들이 결제 커맨드(예를 들어, "$48.12를 결제")을 사용하여 그들의 음성으로 결제하는 것일 수 있다.

[0125] 선택적으로, 스피치 대 스피치 변환 기술의 잠재적 악용에 대처하기 위해, 시스템(100)은 스피치 샘플이 가짜라는 것(즉, 지어낸 것)을 증명하기 위해 쉽게 검출될 수 있는 주파수 성분들("워터마크")을 추가할 수 있다. 이것은, 예를 들어, 인간에게 의해 들리지 않는 저주파 사운드들을 추가하는 것에 의해 달성될 수 있다. 따라서, 워터마크는 인간들에 의해 인식 불가능할 수 있다.

[0126] 본 발명은 전술한 예시적인 실시예들을 통해 설명되었지만, 예시된 실시예들에 대한 수정들 및 그 변형들이 본

명세서에 개시된 발명의 개념들로부터 벗어나지 않고 이루어질 수 있다. 또한, 개시된 양태들 또는 그 부분들은 위에서 열거되지 않고/않았거나 명시적으로 청구되지 않는 방식들로 조합될 수 있다. 따라서, 본 발명은 개시된 실시예들에 제한되는 것으로 간주되어서는 안 된다.

[0127] 본 발명의 다양한 실시예들은 적어도 부분적으로는 임의의 종래의 컴퓨터 프로그래밍 언어로 구현될 수 있다. 예를 들어, 일부 실시예들은 절차적 프로그래밍 언어(예를 들어, "C")로 또는 객체 지향 프로그래밍 언어(예를 들어, "C++")로 구현될 수 있다. 본 발명의 다른 실시예들은 사전 구성된 독립형 하드웨어 요소 및/또는 사전 프로그래밍된 하드웨어 요소들(예를 들어, 주문형 집적 회로들, FPGA들, 및 디지털 신호 프로세서들) 또는 다른 관련 컴포넌트들로서 구현될 수 있다.

[0128] 대안 실시예들에서, 개시된 장치들 및 방법들(예를 들어, 전술한 다양한 흐름도들 참조)은 컴퓨터 시스템에서 사용하기 위한 컴퓨터 프로그램 제품으로서 구현될 수 있다. 이러한 구현은 컴퓨터 판독가능 매체(예를 들어, 디스켓, CD-ROM, ROM 또는 고정 디스크)와 같은 유형적인 비일시적 매체 상에 고정된 일련의 컴퓨터 명령어들을 포함할 수 있다. 일련의 컴퓨터 명령어들은 시스템과 관련하여 본 명세서에서 전술한 기능의 전부 또는 일부를 구현할 수 있다.

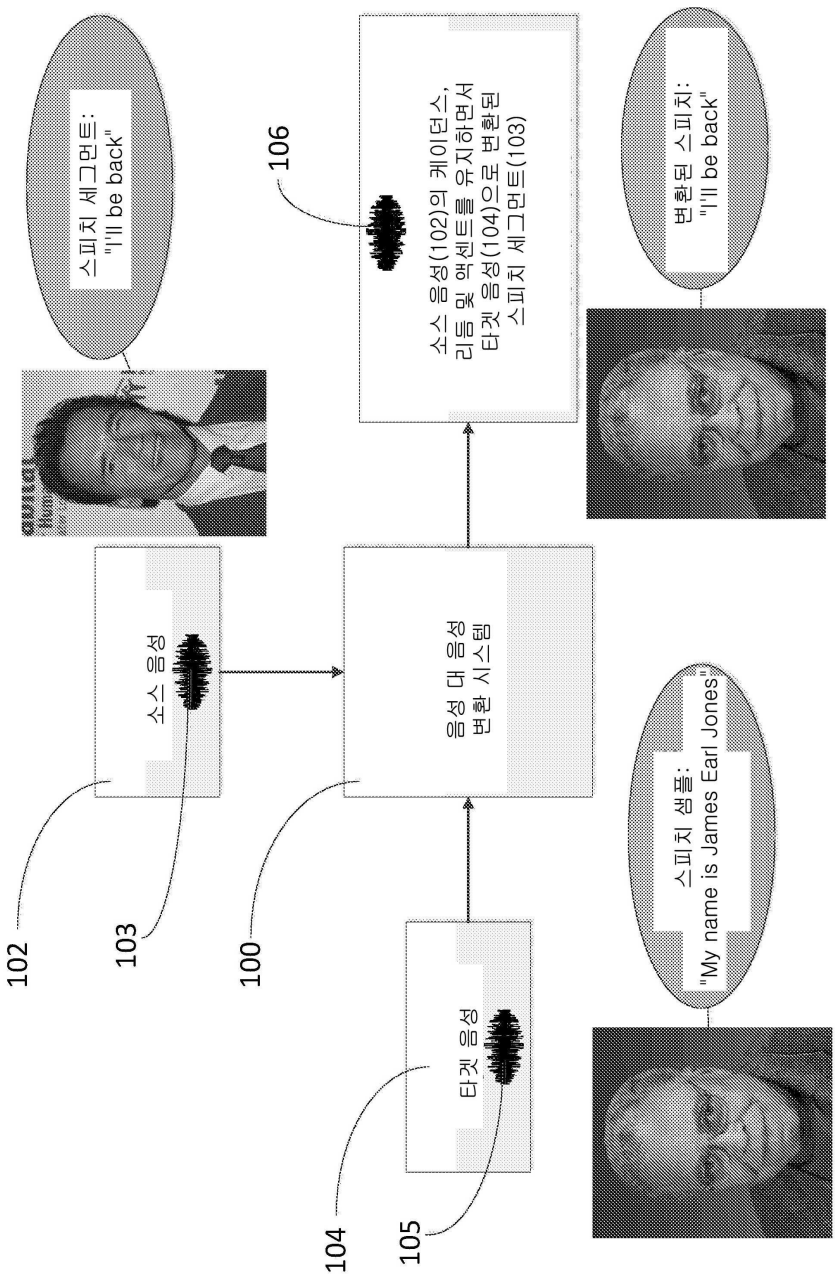
[0129] 이 분야의 통상의 기술자들은 이러한 컴퓨터 명령어들이 많은 컴퓨터 아키텍처 또는 운영 체제와 함께 사용되도록 다수의 프로그래밍 언어로 작성될 수 있다는 것을 알아야 한다. 또한, 이러한 명령어들은 반도체, 자기, 광 또는 다른 메모리 디바이스들과 같은 임의의 메모리 디바이스에 저장될 수 있고, 광, 적외선, 마이크로웨이브 또는 다른 전송 기술들과 같은 임의의 통신 기술을 사용하여 전송될 수 있다.

[0130] 많은 방식 가운데 특히, 이러한 컴퓨터 프로그램 제품은 동반하는 인쇄된 또는 전자적 문서(예컨대, 축소 포장된 소프트웨어)와 함께 이동식 매체로서 배포되거나, 컴퓨터 시스템(예컨대, 시스템 ROM 또는 고정 디스크)에 사전 로딩되거나, 네트워크(예컨대, 인터넷 또는 월드 와이드 웹)을 통해 서버 또는 전자 게시판으로부터 배포될 수 있다. 실제로, 일부 실시예들은 "SAAS" 모델(software-as-a-service model) 또는 클라우드 컴퓨팅 모델에서 구현될 수 있다. 물론, 본 발명의 일부 실시예들은 소프트웨어(예를 들어, 컴퓨터 프로그램 제품) 및 하드웨어의 조합으로서 구현될 수 있다. 본 발명의 또 다른 실시예들은 완전히 하드웨어로서 또는 완전히 소프트웨어로서 구현된다.

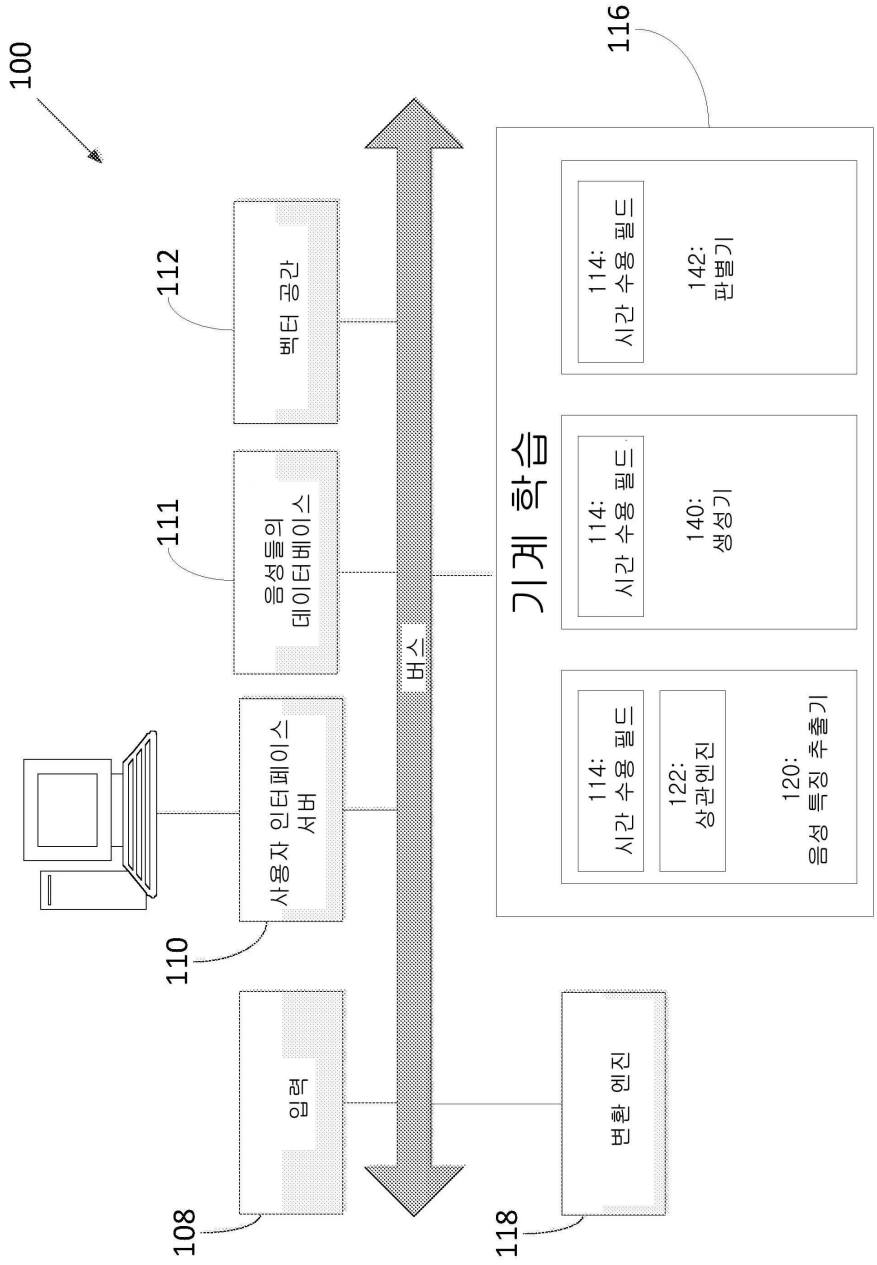
[0131] 전술한 본 발명의 실시예들은 단지 예시적인 것으로 의도되며; 이 분야의 통상의 기술자들에게는 많은 변경과 수정이 자명할 것이다. 그러한 변경들 및 수정들은 첨부된 청구항들 중 임의의 청구항에 의해 정의되는 바와 같은 본 발명의 범위 내에 있는 것으로 의도된다.

도면

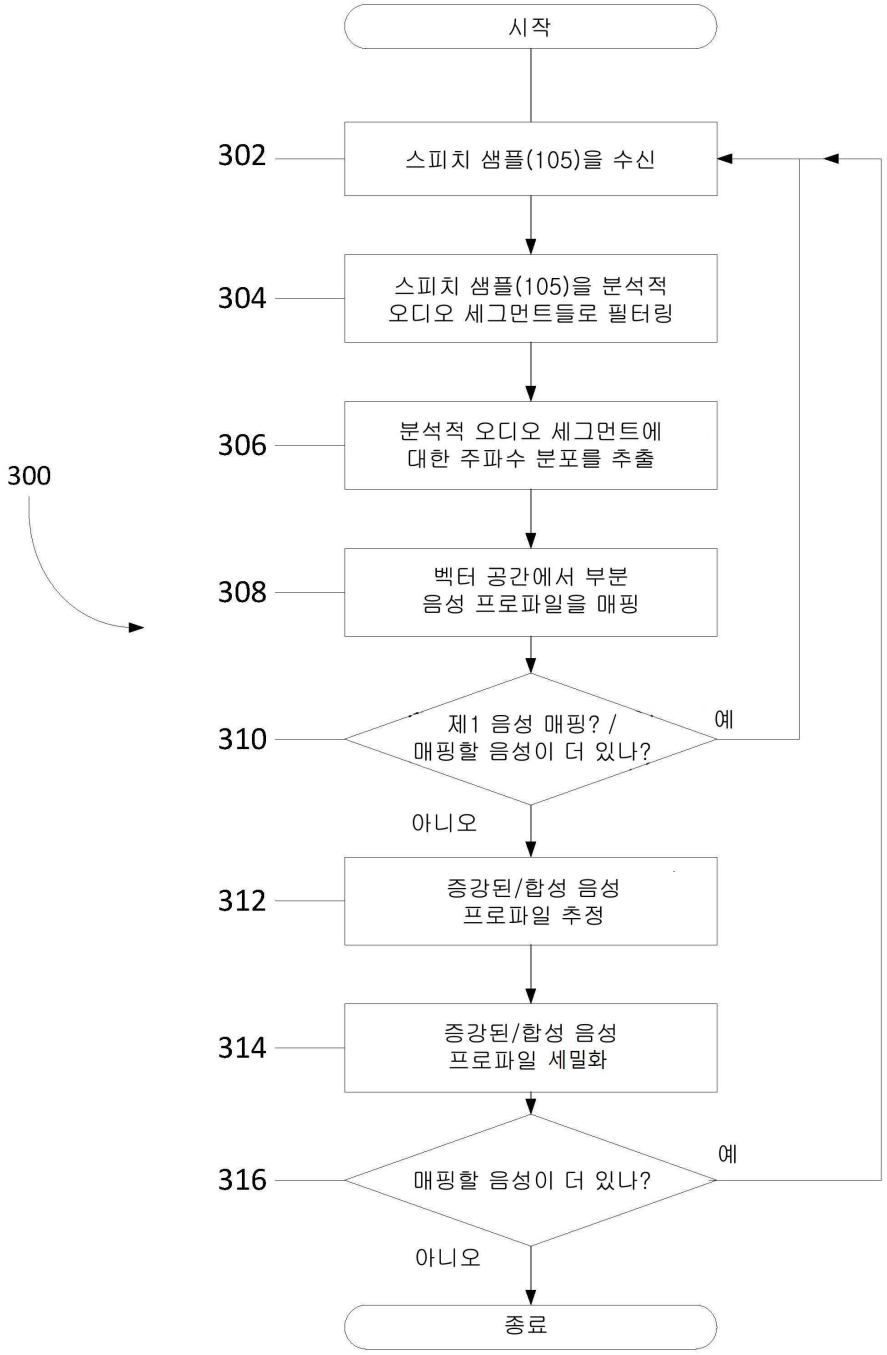
도면1


스피치 세그먼트:
"I'll be back"
106
소스 음성(102)의 케이던스,
리듬 및 액센트를 유지하면서
타겟 음성(104)으로 변환된
스피치 세그먼트(103)
변환된 스피치:
"I'll be back"
소스 음성
음성 대 음성
변환 시스템
스피치 샘플:
"My name is James Earl Jones"
타겟 음성
102
103
100
104
105

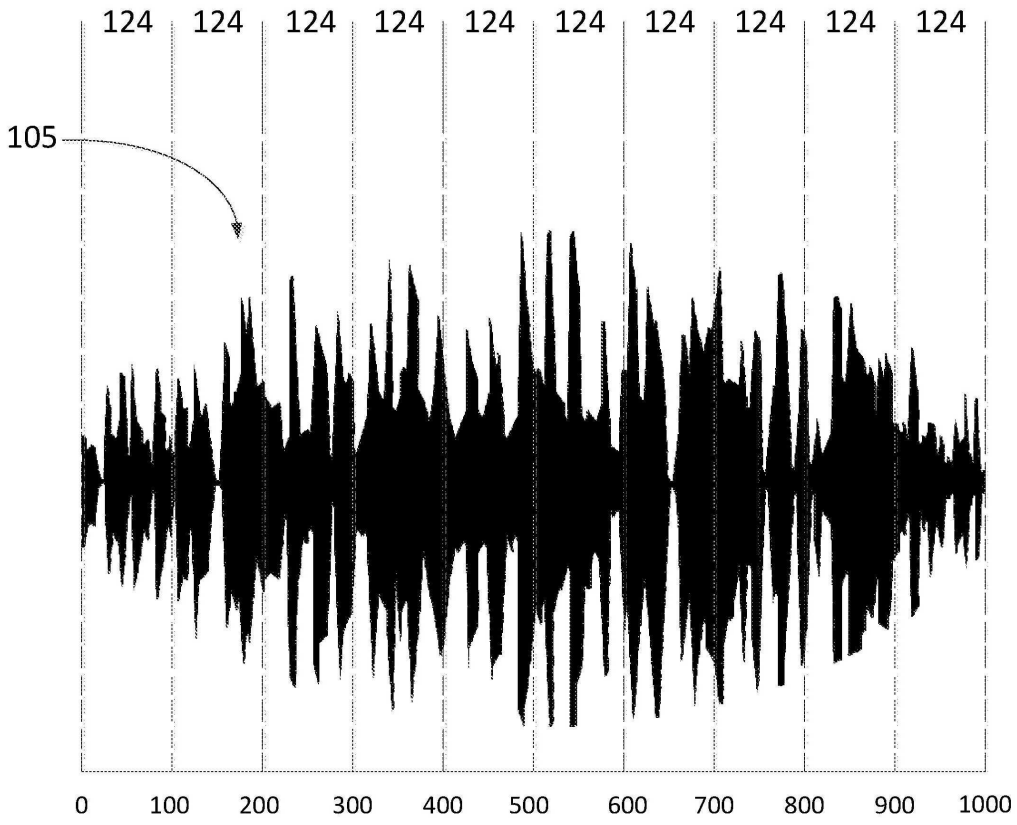
도면2


100
116
112
111
110
108
118
벡터 공간
음성들의 데이터베이스
사용자 인터페이스 서버
입력
변환 엔진
버스
기계 학습
114: 시간 수용 필드
142: 판별기
114: 시간 수용 필드
140: 생성기
114: 시간 수용 필드
122: 상관엔진
120: 음성 특징 추출기

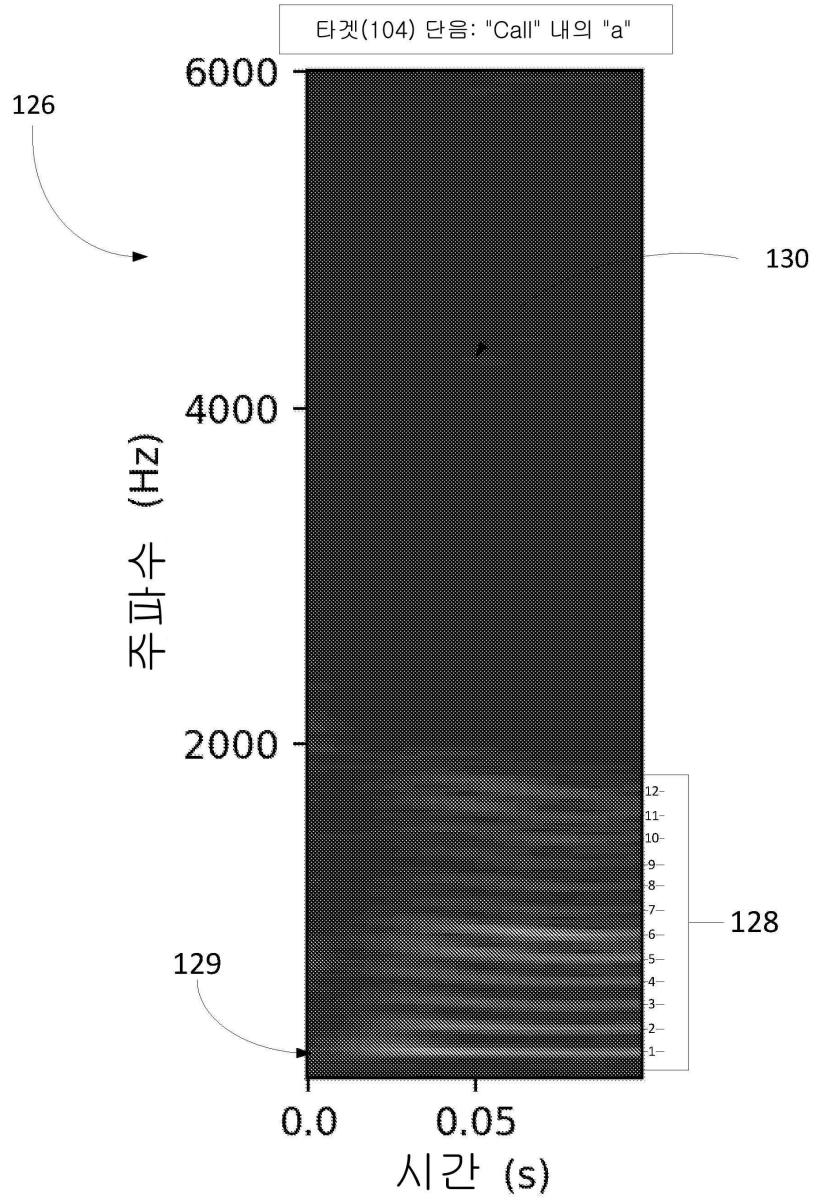
도면3


시작
302
스피치 샘플(105)을 수신
304
스피치 샘플(105)을 분석적 오디오 세그먼트들로 필터링
306
분석적 오디오 세그먼트에 대한 주파수 분포를 추출
300
308
벡터 공간에서 부분 음성 프로파일을 매핑
310
제1 음성 매핑? / 매핑할 음성이 더 있나?
예
아니오
312
증강된/합성 음성 프로파일 추정
314
증강된/합성 음성 프로파일 세밀화
316
매핑할 음성이 더 있나?
예
아니오
종료

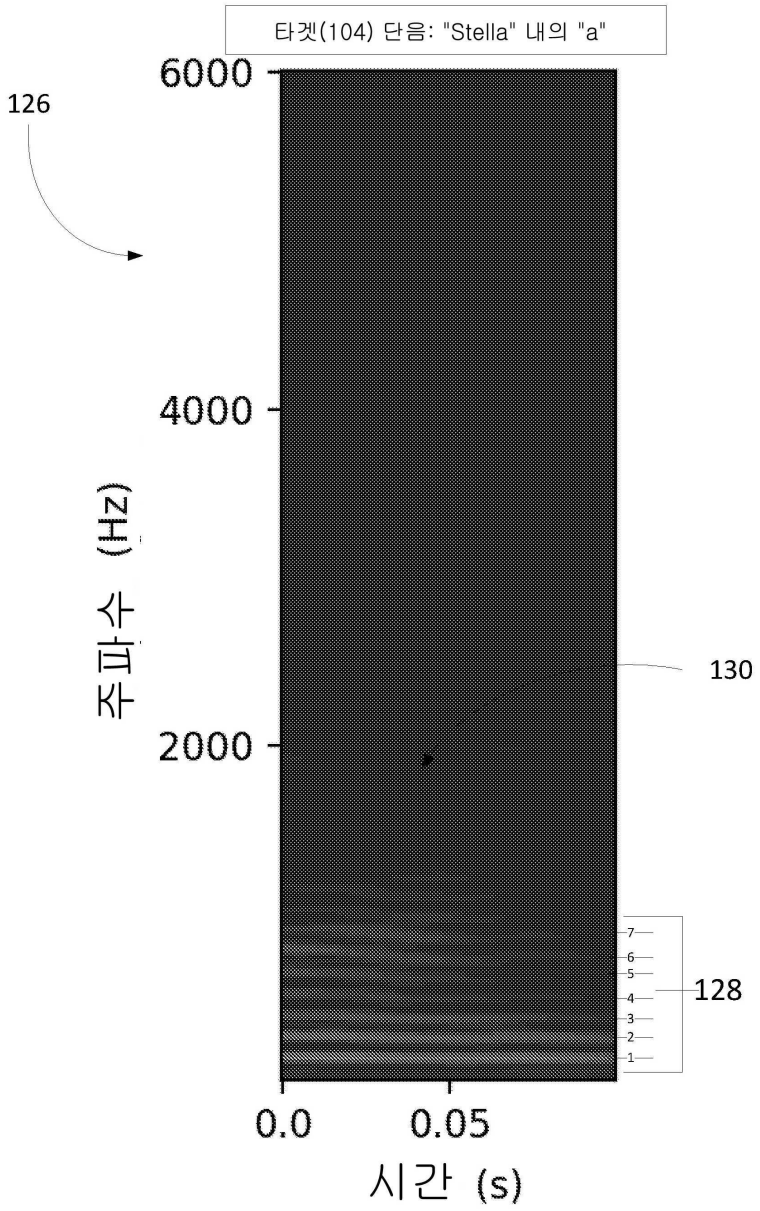
도면4


124
124
124
124
124
124
124
124
124
124
105
0
100
200
300
400
500
600
700
800
900
1000
시간(밀리초)

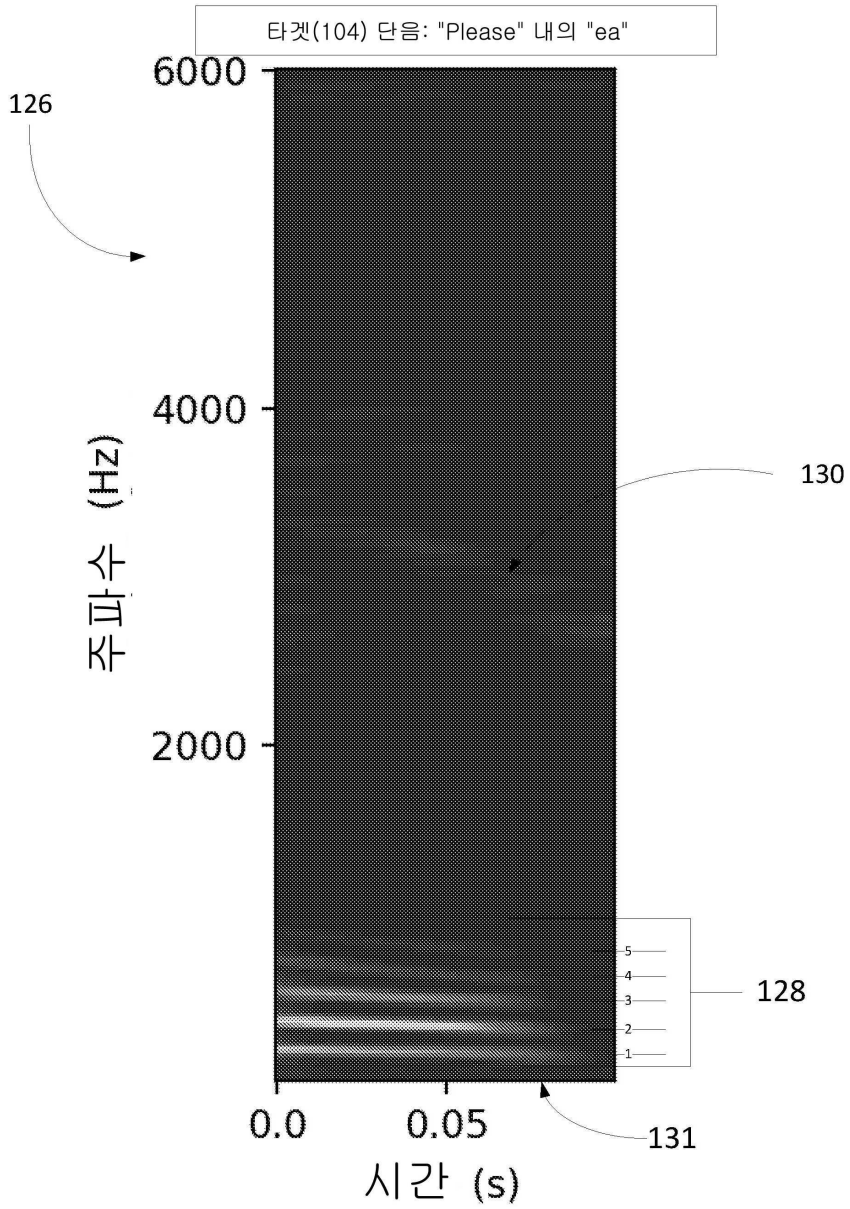
도면5a


타겟(104) 단음: "Call" 내의 "a"
126
130
6000
4000
2000
주파수 (Hz)
128
129
0.0
0.05
시간 (s)

도면5b

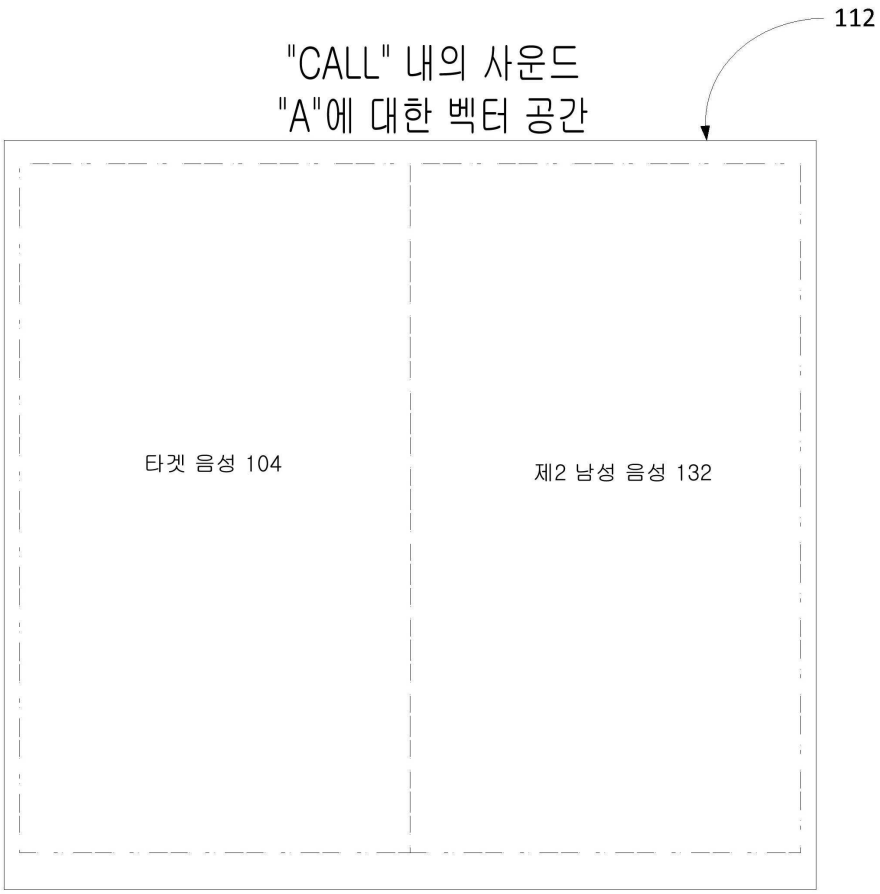

타겟(104) 단음: "Stella" 내의 "a"
126
6000
4000
2000
주파수 (Hz)
130
128
7
6
5
4
3
2
1
0.0
0.05
시간 (s)

도면5c

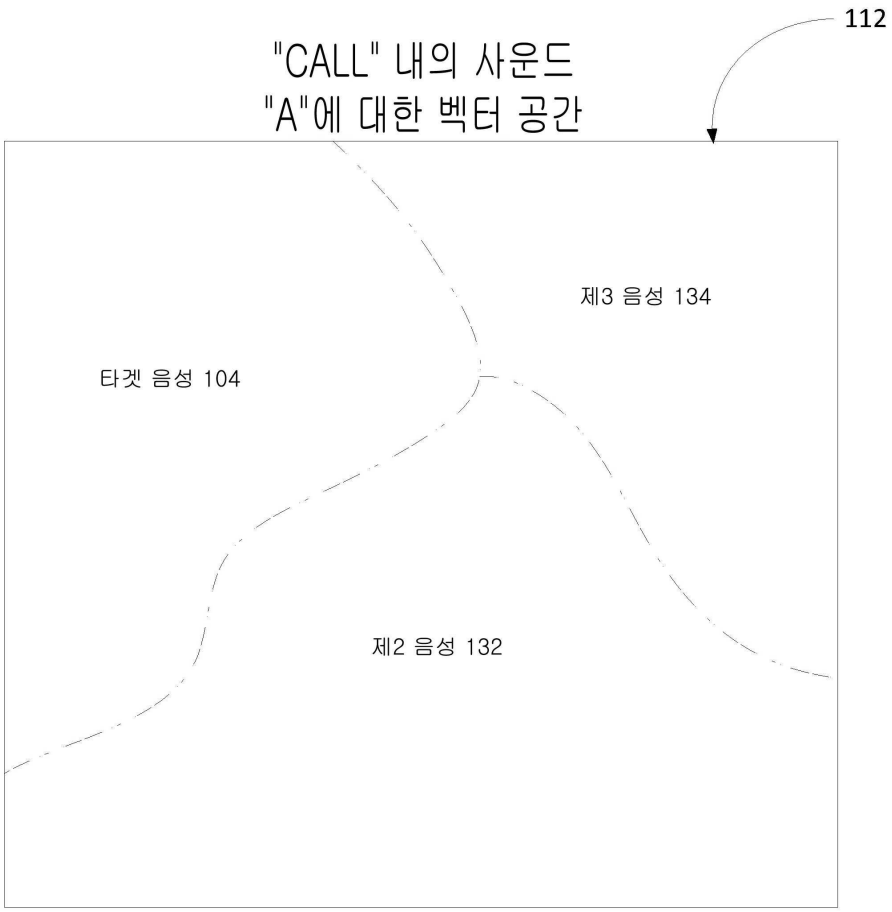

타겟(104) 단음: "Please" 내의 "ea"
126
6000
4000
2000
주파수 (Hz)
130
128
5
4
3
2
1
0.0
0.05
131
시간 (s)

도면6a

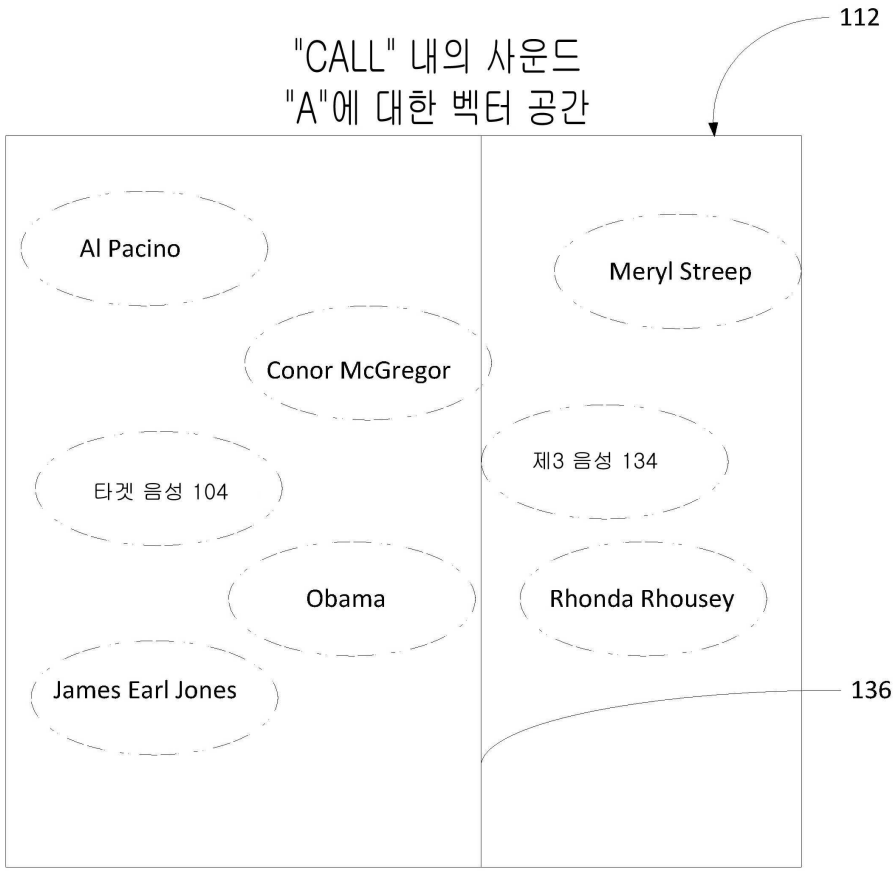
도면6b



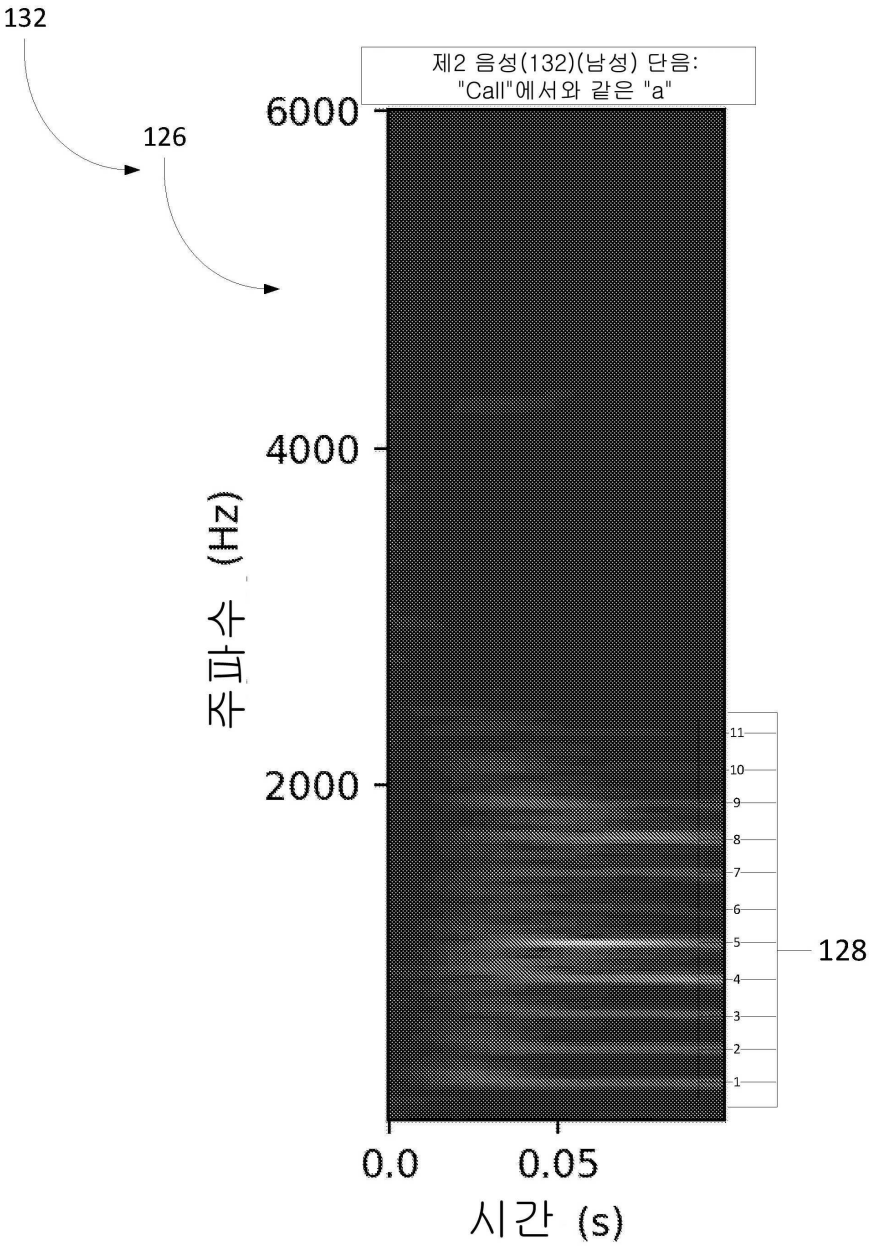
도면6c



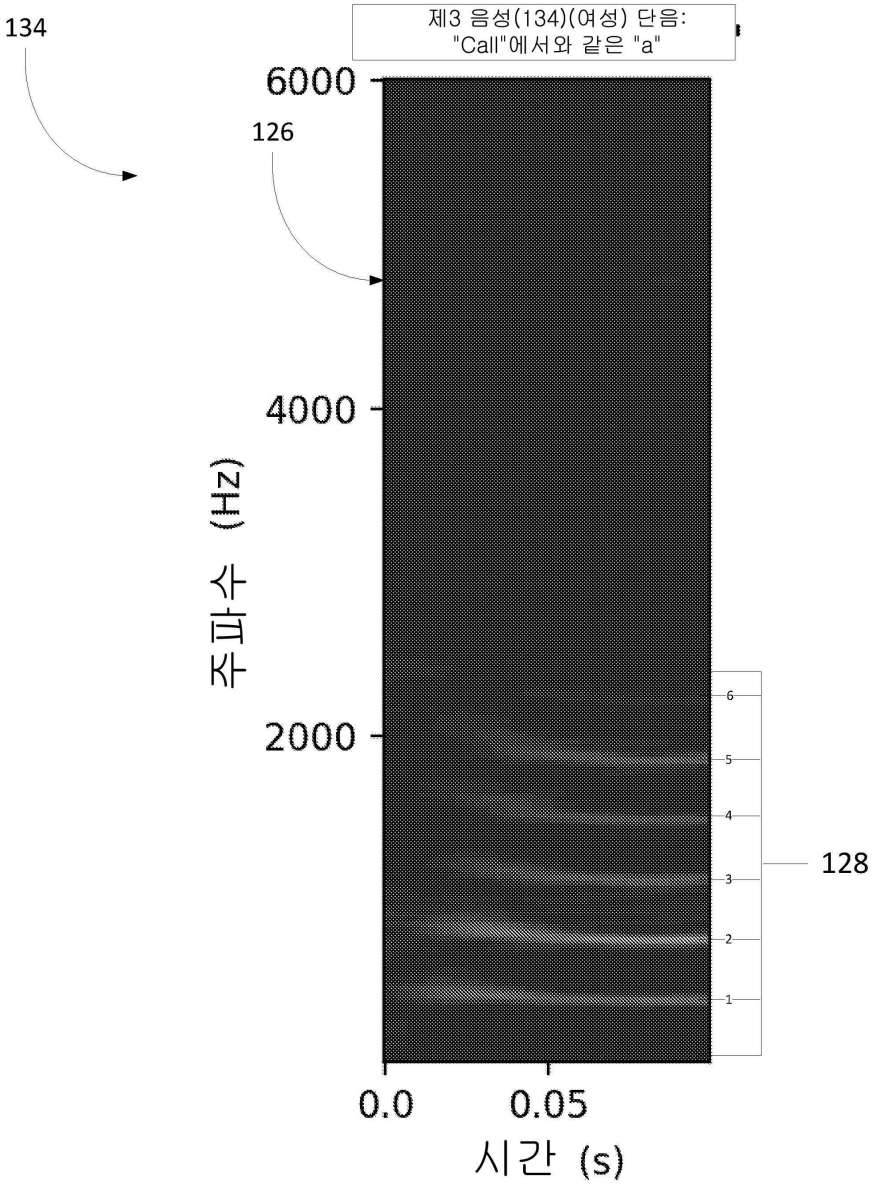
도면6d


"CALL" 내의 사운드
"A"에 대한 벡터 공간
112
Al Pacino
Meryl Streep
Conor McGregor
제3 음성 134
타겟 음성 104
Obama
Rhonda Rhousey
James Earl Jones
136

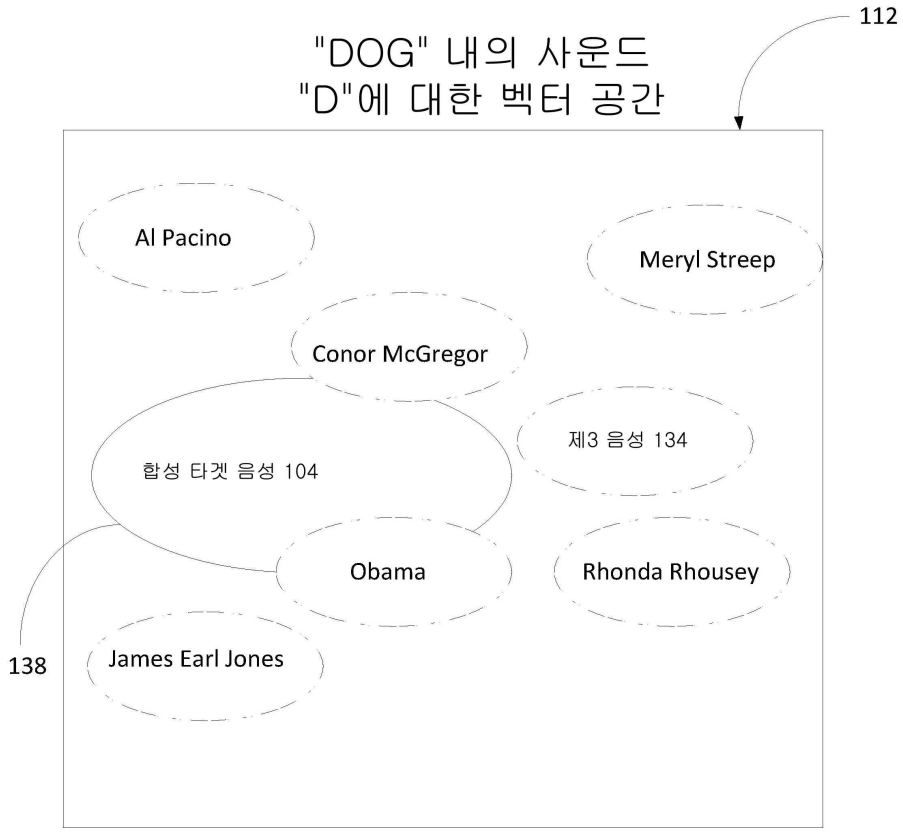
도면7a


132
126
제2 음성(132)(남성) 단음:
"Call"에서와 같은 "a"
6000
4000
2000
주파수 (Hz)
11
10
9
8
7
6
5
4
3
2
1
128
0.0
0.05
시간 (s)

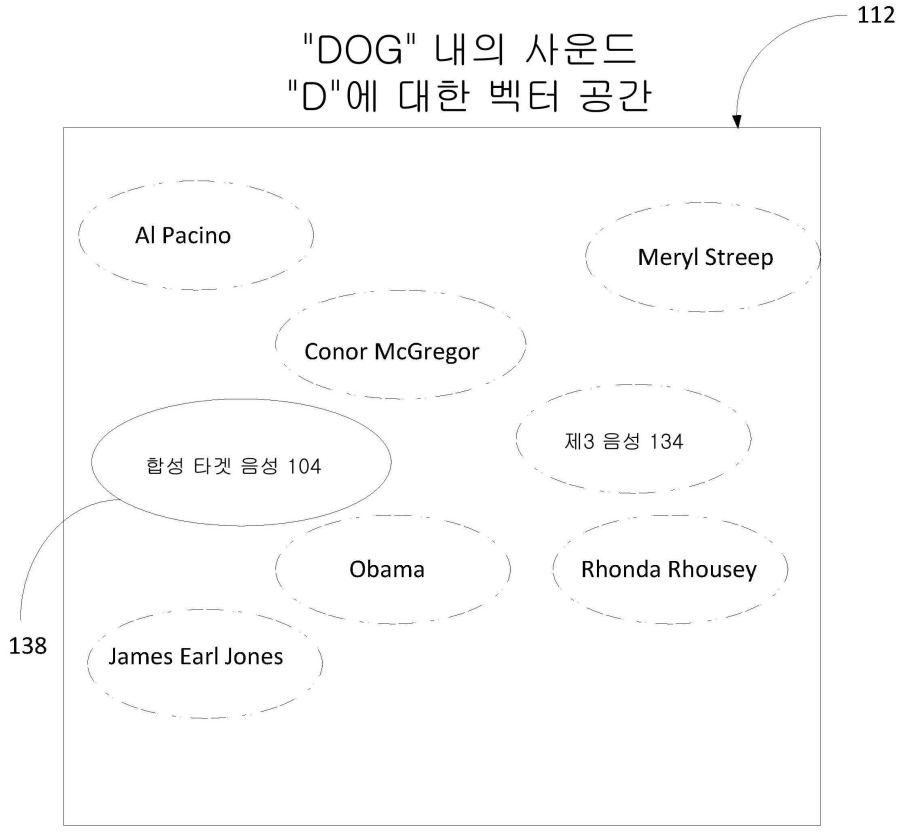
도면7b


134
제3 음성(134)(여성) 단음:
"Call"에서와 같은 "a"
6000
126
4000
2000
주파수 (Hz)
6
5
4
3
2
1
128
0.0
0.05
시간 (s)

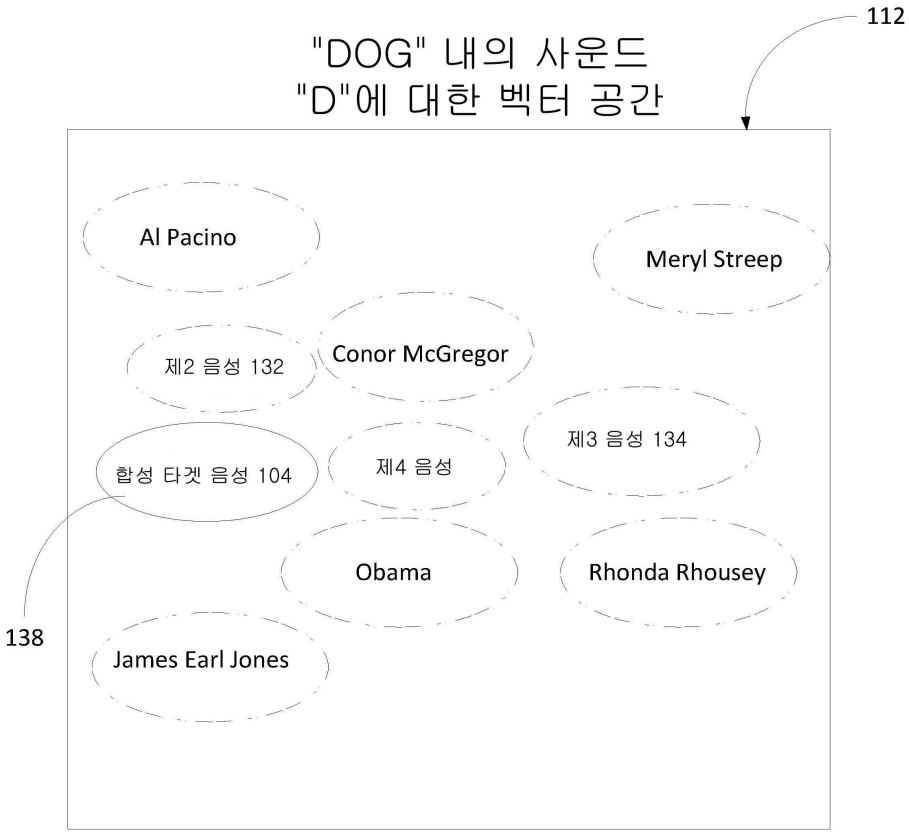
도면8a


"DOG" 내의 사운드
"D"에 대한 벡터 공간
112
Al Pacino
Meryl Streep
Conor McGregor
제3 음성 134
합성 타겟 음성 104
Obama
Rhonda Rhousey
138
James Earl Jones

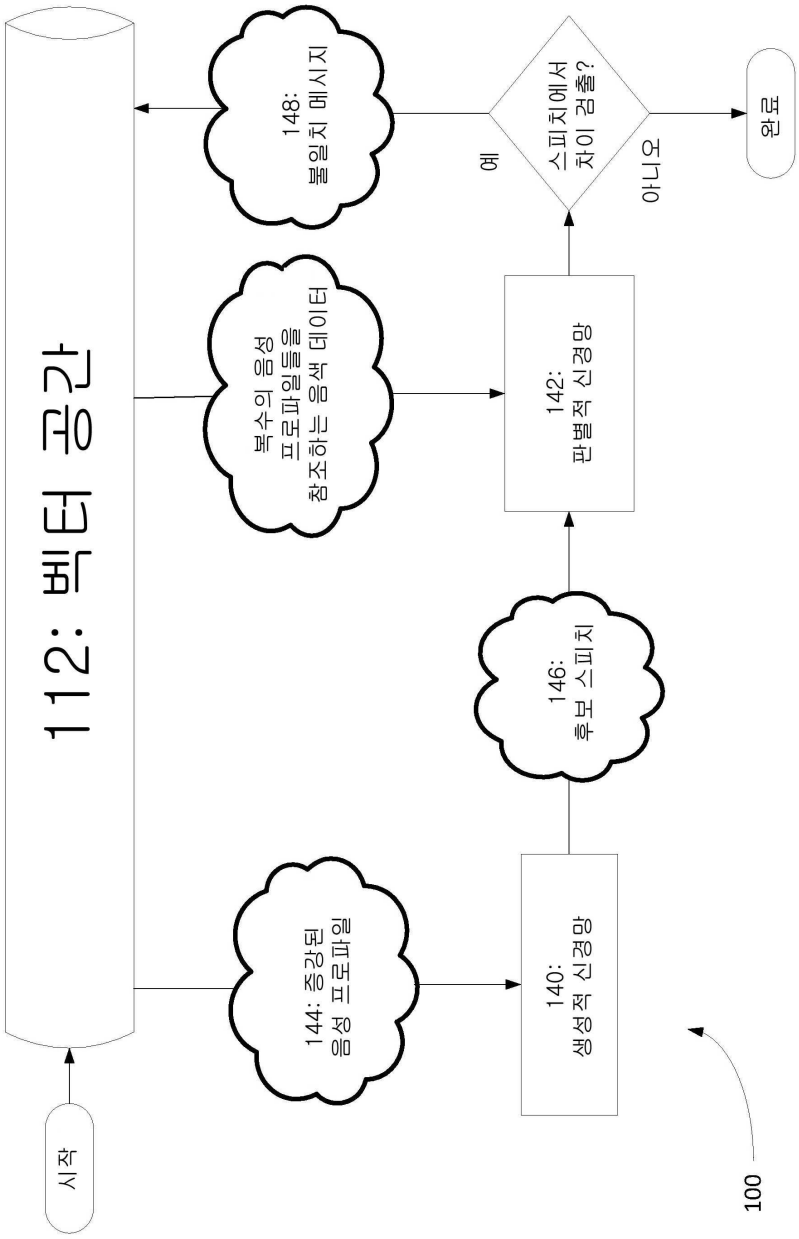
도면8b


"DOG" 내의 사운드
"D"에 대한 벡터 공간
112
Al Pacino
Meryl Streep
Conor McGregor
제3 음성 134
합성 타겟 음성 104
Obama
Rhonda Rhousey
138
James Earl Jones

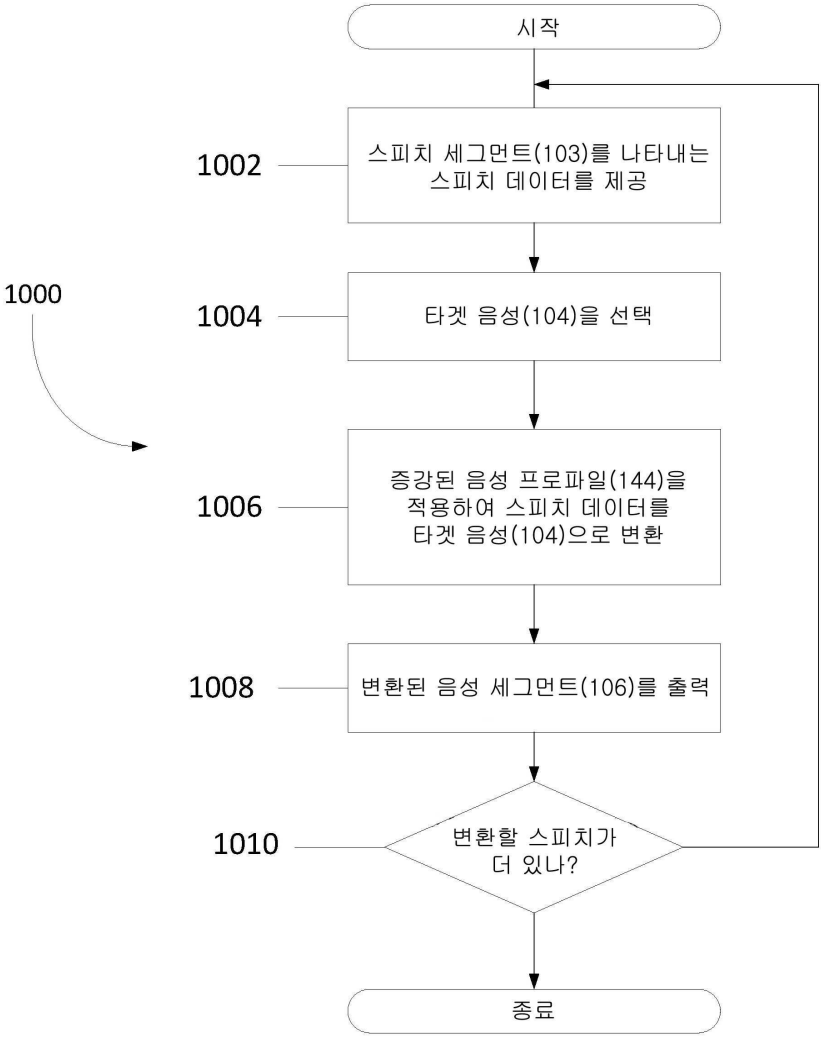
도면8c


"DOG" 내의 사운드
"D"에 대한 벡터 공간
112
Al Pacino
Meryl Streep
제2 음성 132
Conor McGregor
제3 음성 134
합성 타겟 음성 104
제4 음성
Obama
Rhonda Rhousey
James Earl Jones
138

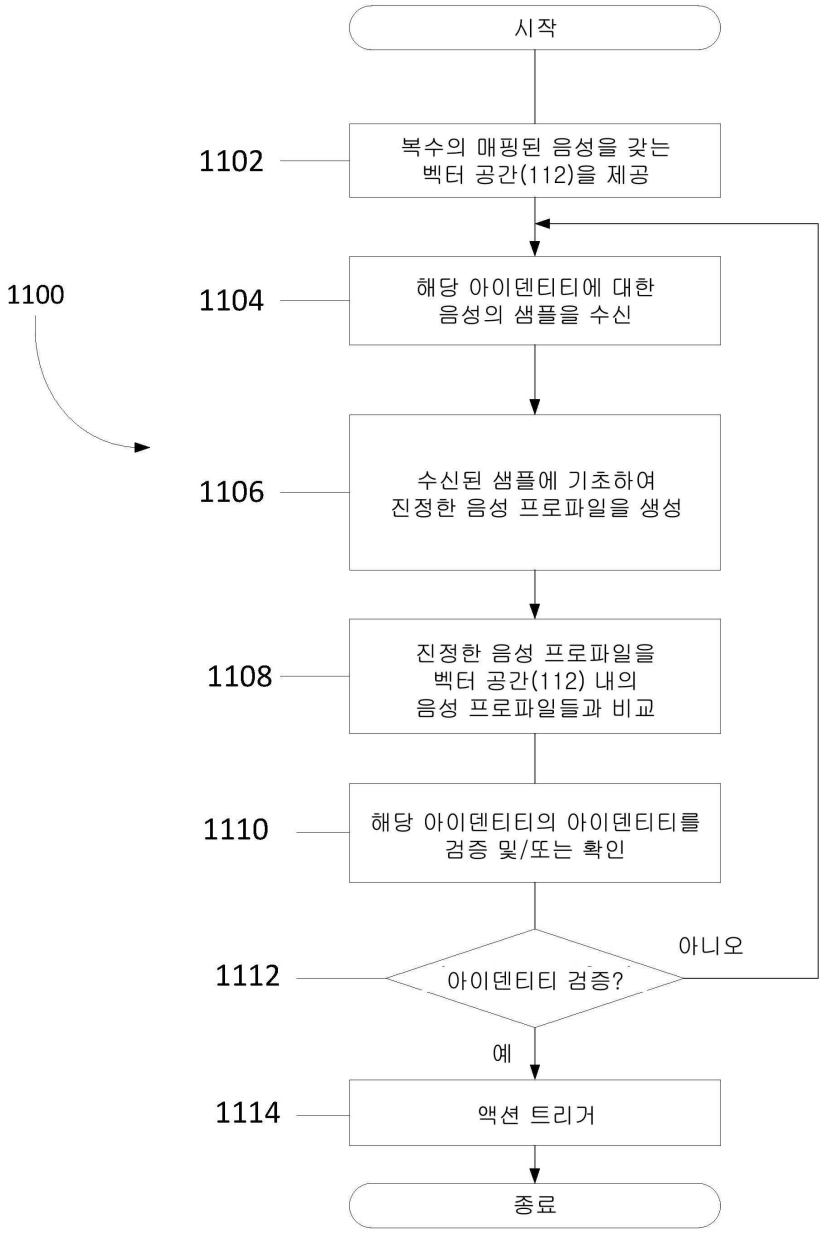
도면9


시작
112: 벡터 공간
144: 증강된 음성 프로파일
140: 생성적 신경망
146: 후보 스피치
복수의 음성 프로파일들을 참조하는 음색 데이터
142: 판별적 신경망
148: 불일치 메시지
스피치에서 차이 검출?
예
아니오
완료
100

도면10


시작
1002
스피치 세그먼트(103)를 나타내는 스피치 데이터를 제공
1000
1004
타겟 음성(104)을 선택
1006
증강된 음성 프로파일(144)을 적용하여 스피치 데이터를 타겟 음성(104)으로 변환
1008
변환된 음성 세그먼트(106)를 출력
1010
변환할 스피치가 더 있나?
종료

도면11


시작
1102
복수의 매핑된 음성을 갖는 벡터 공간(112)을 제공
1100
1104
해당 아이덴티티에 대한 음성의 샘플을 수신
1106
수신된 샘플에 기초하여 진정한 음성 프로파일을 생성
1108
진정한 음성 프로파일을 벡터 공간(112) 내의 음성 프로파일들과 비교
1110
해당 아이덴티티의 아이덴티티를 검증 및/또는 확인
1112
아이덴티티 검증?
아니오
예
1114
액션 트리거
종료