



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0133295
(43) 공개일자 2023년09월19일

(51) 국제특허분류(Int. Cl.)
G10L 13/047 (2013.01) G06N 3/045 (2023.01)
G10L 13/033 (2013.01) G10L 19/02 (2006.01)
G10L 21/007 (2013.01) G10L 21/013 (2013.01)
G10L 25/63 (2013.01)
(52) CPC특허분류
G10L 13/047 (2013.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2023-7024214
(22) 출원일자(국제) 2021년12월08일
심사청구일자 없음
(85) 번역문제출일자 2023년07월14일
(86) 국제출원번호 PCT/US2021/072800
(87) 국제공개번호 WO 2022/159256
국제공개일자 2022년07월28일
(30) 우선권주장
17/154,372 2021년01월21일 미국(US)

(71) 출원인
켈컴 인코포레이티드
미국 92121-1714 캘리포니아주 샌 디에고 모어하우스 드라이브 5775
(72) 발명자
변 경근
미국 92121 캘리포니아주 샌디에고 모어하우스 드라이브 5775
문 선국
미국 92121 캘리포니아주 샌디에고 모어하우스 드라이브 5775
(74) 대리인
(뒷면에 계속)
특허법인코리아나

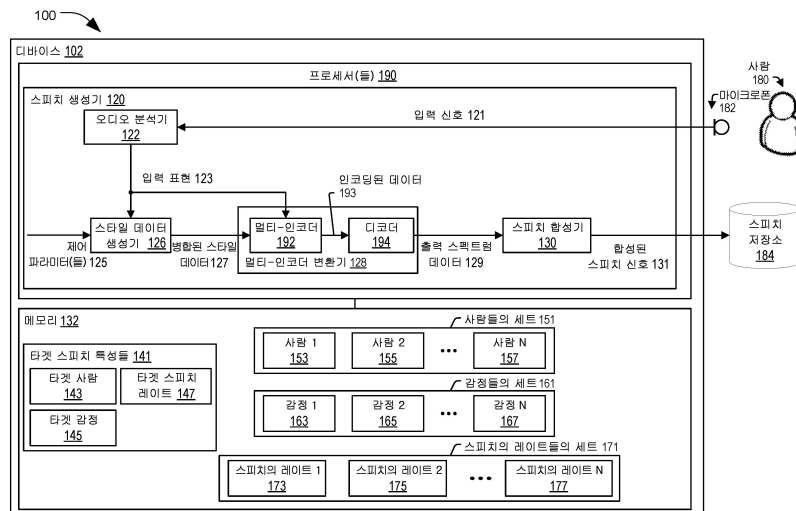
전체 청구항 수 : 총 30 항

(54) 발명의 명칭 합성된 스피치 생성

(57) 요약

스피치 생성을 위한 디바이스는 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하도록 구성된 하나 이상의 프로세서들을 포함한다. 하나 이상의 프로세서들은 또한, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하도록 구성된다.

대표도



(52) CPC특허분류

G10L 13/033 (2013.01)

G10L 19/02 (2013.01)

G10L 21/007 (2013.01)

G10L 25/63 (2013.01)

G10L 2021/0135 (2013.01)

(72) 발명자

장 슈화

미국 92121 캘리포니아주 샌디에고 모어하우스 드
라이브 5775

몬타제리 바히드

미국 92121 캘리포니아주 샌디에고 모어하우스 드
라이브 5775

김 래훈

미국 92121 캘리포니아주 샌디에고 모어하우스 드
라이브 5775

비제르 에릭

미국 92121 캘리포니아주 샌디에고 모어하우스 드
라이브 5775

명세서

청구범위

청구항 1

스피치 생성을 위한 디바이스로서,

하나 이상의 프로세서들을 포함하고,

상기 하나 이상의 프로세서들은,

타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하고; 그리고

상기 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들에 기초하여 상기 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하도록 구성되는, 스피치 생성을 위한 디바이스.

청구항 2

제 1 항에 있어서,

상기 제어 파라미터들은, 스피치 특성들이 사용될 타겟 사람, 타겟 감정, 타겟 스피치 레이트, 또는 이들의 조합을 표시하는, 스피치 생성을 위한 디바이스.

청구항 3

제 1 항에 있어서,

상기 하나 이상의 프로세서들은 추가로, 상기 하나 이상의 제어 파라미터들에 기초하여 병합된 스타일 데이터를 생성하도록 구성되고, 상기 병합된 스타일 데이터는 상기 입력 표현의 프로세싱 동안 상기 멀티-인코더에 의해 사용되는, 스피치 생성을 위한 디바이스.

청구항 4

제 1 항에 있어서,

상기 멀티-인코더는,

제 1 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들과 독립적으로 상기 입력 표현을 인코딩하도록 구성된 제 1 인코더; 및

제 2 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들에 기초하여 상기 입력 표현을 인코딩하도록 구성된 하나 이상의 제 2 인코더들을 포함하고,

상기 인코딩된 데이터는 상기 제 1 인코딩된 데이터 및 상기 제 2 인코딩된 데이터를 포함하는, 스피치 생성을 위한 디바이스.

청구항 5

제 4 항에 있어서,

상기 하나 이상의 프로세서들은 추가로,

스피치 특성 인코더에서, 인코딩된 입력 스피치 표현을 생성하기 위해 상기 하나 이상의 제어 파라미터들 중 적어도 하나에 기초하여 상기 입력 표현을 프로세싱하고;

인코더 프리-네트워크에서, 상기 인코딩된 입력 스피치 표현에 적어도 부분적으로 기초하여 병합된 스타일 데이터를 생성하고;

상기 제 1 인코딩된 데이터를 생성하기 위해 상기 제 1 인코더에 상기 입력 표현을 제공하며; 그리고

상기 제 2 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제 2 인코더들에 상기 병합된 스타일 데이터를 제공하도록 구성되는, 스피치 생성을 위한 디바이스.

청구항 6

제 4 항에 있어서,

상기 멀티-인코더 및 디코더를 포함하는 멀티-인코더 변환기를 더 포함하고, 상기 제 1 인코더는 제 1 어텐션 네트워크를 포함하고, 상기 하나 이상의 제 2 인코더들의 각각은 제 2 어텐션 네트워크를 포함하며, 상기 디코더는 디코더 어텐션 네트워크를 포함하는, 스피치 생성을 위한 디바이스.

청구항 7

제 6 항에 있어서,

상기 제 1 인코더는,

상기 제 1 어텐션 네트워크를 포함하는 제 1 계층으로서, 상기 제 1 어텐션 네트워크는 제 1 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 1 어텐션 네트워크를 포함하는 제 1 계층; 및

제 1 뉴럴 네트워크를 포함하는 제 2 계층을 포함하고,

상기 하나 이상의 제 2 인코더들의 각각은,

상기 제 2 어텐션 네트워크를 포함하는 제 1 계층으로서, 상기 제 2 어텐션 네트워크는 제 2 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 2 어텐션 네트워크를 포함하는 제 1 계층; 및

제 2 뉴럴 네트워크를 포함하는 제 2 계층을 포함하는, 스피치 생성을 위한 디바이스.

청구항 8

제 4 항에 있어서,

상기 멀티-인코더에 커플링된 디코더로서, 상기 디코더는 상기 제 1 인코딩된 데이터 및 상기 제 2 인코딩된 데이터에 기초하여 출력 스펙트럼 데이터를 생성하도록 구성되는 디코더 네트워크를 포함하는, 상기 디코더; 및

상기 타겟 스피치 특성들에 기초하여 상기 스피치의 버전을 표현하는 상기 오디오 신호를, 상기 출력 스펙트럼 데이터에 기초하여, 생성하도록 구성된 스피치 합성기를 더 포함하는, 스피치 생성을 위한 디바이스.

청구항 9

제 8 항에 있어서,

상기 디코더 네트워크는 디코더 어텐션 네트워크를 포함하고, 상기 디코더 어텐션 네트워크는,

상기 제 1 인코딩된 데이터를 프로세싱하도록 구성된 제 1 멀티-헤드 어텐션 네트워크;

상기 제 2 인코딩된 데이터를 프로세싱하도록 구성된 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들; 및

상기 제 1 멀티-헤드 어텐션 네트워크 및 상기 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들의 출력들을 결합하도록 구성된 결합기를 포함하는, 스피치 생성을 위한 디바이스.

청구항 10

제 9 항에 있어서,

상기 디코더는,

상기 디코더 어텐션 네트워크의 입력에 커플링된 마스킹된 멀티-헤드 어텐션 네트워크; 및

상기 디코더 어텐션 네트워크의 출력에 커플링된 디코더 뉴럴 네트워크를 더 포함하는, 스피치 생성을 위한 디바이스.

청구항 11

제 1 항에 있어서,

상기 하나 이상의 프로세서들은 추가로,

상기 오디오 신호로부터 하나 이상의 추정된 제어 파라미터들을 생성하고; 그리고

상기 하나 이상의 제어 파라미터들과 상기 하나 이상의 추정된 제어 파라미터들의 비교에 기초하여, 상기 멀티-인코더, 하나 이상의 스피치 특성 인코더들, 인코더 프리-네트워크, 디코더 네트워크, 또는 이들의 조합의 하나 이상의 뉴럴 네트워크 가중치들을 훈련하도록 구성되는, 스피치 생성을 위한 디바이스.

청구항 12

제 1 항에 있어서,

상기 하나 이상의 프로세서들은 추가로,

입력 스피치 신호를 수신하고; 그리고

상기 입력 스피치 신호에 기초하여 상기 입력 표현을 생성하도록 구성되는, 스피치 생성을 위한 디바이스.

청구항 13

제 1 항에 있어서,

상기 하나 이상의 프로세서들은 추가로, 상기 입력 표현을 수신하도록 구성되는, 스피치 생성을 위한 디바이스.

청구항 14

제 1 항에 있어서,

상기 입력 표현은 텍스트, 멜-스케일(mel-scale) 스펙트로그램들, 기본 주파수 (F0) 특징들, 또는 이들의 조합을 포함하는, 스피치 생성을 위한 디바이스.

청구항 15

스피치 생성 방법으로서,

디바이스에서, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 단계; 및

상기 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들에 기초하여 상기 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 단계를 포함하는, 스피치 생성 방법.

청구항 16

제 15 항에 있어서,

상기 제어 파라미터들은, 스피치 특성들이 사용될 타겟 사람, 타겟 감정, 타겟 스피치 레이트, 또는 이들의 표시하는, 스피치 생성 방법.

청구항 17

제 15 항에 있어서,

상기 디바이스에서, 상기 하나 이상의 제어 파라미터들에 기초하여 병합된 스타일 데이터를 생성하는 단계를 더 포함하고, 상기 병합된 스타일 데이터는 상기 입력 표현의 프로세싱 동안 상기 멀티-인코더에 의해 사용되는, 스피치 생성 방법.

청구항 18

제 15 항에 있어서,

상기 멀티-인코더의 제 1 인코더에서, 제 1 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들과 독립적으로 상기 입력 표현을 인코딩하는 단계; 및

상기 멀티-인코더의 하나 이상의 제 2 인코더들에서, 제 2 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들에 기초하여 상기 입력 표현을 인코딩하는 단계를 더 포함하고,

상기 인코딩된 데이터는 상기 제 1 인코딩된 데이터 및 상기 제 2 인코딩된 데이터를 포함하는, 스피치 생성 방법.

청구항 19

제 18 항에 있어서,

스피치 특성 인코더에서, 인코딩된 입력 스피치 표현을 생성하기 위해 상기 하나 이상의 제어 파라미터들 중 적어도 하나에 기초하여 상기 입력 표현을 프로세싱하는 단계;

인코더 프리-네트워크에서, 상기 인코딩된 입력 스피치 표현에 적어도 부분적으로 기초하여 병합된 스타일 데이터를 생성하는 단계;

상기 제 1 인코딩된 데이터를 생성하기 위해 상기 제 1 인코더에 상기 입력 표현을 제공하는 단계; 및

상기 제 2 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제 2 인코더들에 상기 병합된 스타일 데이터를 제공하는 단계를 더 포함하는, 스피치 생성 방법.

청구항 20

제 18 항에 있어서,

상기 멀티-인코더 및 디코더가 멀티-인코더 변환기에 포함되고, 상기 제 1 인코더는 제 1 어텐션 네트워크를 포함하고, 상기 하나 이상의 제 2 인코더들의 각각은 제 2 어텐션 네트워크를 포함하며, 상기 디코더는 디코더 어텐션 네트워크를 포함하는, 스피치 생성 방법.

청구항 21

제 20 항에 있어서,

상기 제 1 인코더는,

상기 제 1 어텐션 네트워크를 포함하는 제 1 계층으로서, 상기 제 1 어텐션 네트워크는 제 1 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 1 어텐션 네트워크를 포함하는 제 1 계층; 및

제 1 뉴럴 네트워크를 포함하는 제 2 계층을 포함하고,

상기 하나 이상의 제 2 인코더들의 각각은,

상기 제 2 어텐션 네트워크를 포함하는 제 1 계층으로서, 상기 제 2 어텐션 네트워크는 제 2 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 2 어텐션 네트워크를 포함하는 제 1 계층; 및

제 2 뉴럴 네트워크를 포함하는 제 2 계층을 포함하는, 스피치 생성 방법.

청구항 22

제 18 항에 있어서,

디코더의 디코더 네트워크에서, 상기 제 1 인코딩된 데이터 및 상기 제 2 인코딩된 데이터에 기초하여 출력 스펙트럼 데이터를 생성하는 단계; 및

스피치 합성기에서 그리고 상기 출력 스펙트럼 데이터에 기초하여, 상기 타겟 스피치 특성들에 기초하여 상기 스피치의 버전을 표현하는 상기 오디오 신호를 생성하는 단계를 더 포함하는, 스피치 생성 방법.

청구항 23

제 22 항에 있어서,

디코더 어텐션 네트워크의 제 1 멀티-헤드 어텐션 네트워크에서 상기 제 1 인코딩된 데이터를 프로세싱하는 단계로서, 상기 디코더 네트워크는 상기 디코더 어텐션 네트워크를 포함하는, 상기 제 1 인코딩된 데이터를 프로

세상하는 단계;

상기 디코더 어텐션 네트워크의 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들에서 상기 제 2 인코딩된 데이터를 프로세싱하는 단계; 및

결합기에서, 상기 제 1 멀티-헤드 어텐션 네트워크 및 상기 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들의 출력들을 결합하는 단계를 더 포함하는, 스피치 생성 방법.

청구항 24

제 15 항에 있어서,

상기 디바이스에서, 상기 오디오 신호로부터 하나 이상의 추정된 제어 파라미터들을 생성하는 단계; 및

상기 하나 이상의 제어 파라미터들과 상기 하나 이상의 추정된 제어 파라미터들의 비교에 기초하여, 상기 멀티-인코더, 하나 이상의 스피치 특성 인코더들, 인코더 프리-네트워크, 디코더 네트워크, 또는 이들의 조합의 하나 이상의 뉴럴 네트워크 가중치들을 훈련하는 단계를 더 포함하는, 스피치 생성 방법.

청구항 25

제 15 항에 있어서,

상기 디바이스에서 입력 스피치 신호를 수신하는 단계; 및

상기 디바이스에서, 상기 입력 스피치 신호에 기초하여 상기 입력 표현을 생성하는 단계를 더 포함하는, 스피치 생성 방법.

청구항 26

제 15 항에 있어서,

상기 디바이스에서 상기 입력 표현을 수신하는 단계를 더 포함하는, 스피치 생성 방법.

청구항 27

명령들을 저장하는 비일시적 컴퓨터 판독가능 저장 매체로서,

상기 명령들은, 하나 이상의 프로세서들에 의해 실행될 때, 상기 하나 이상의 프로세서들로 하여금,

타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하게 하고; 그리고

상기 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들에 기초하여 상기 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하게 하는, 비일시적 컴퓨터 판독가능 저장 매체.

청구항 28

제 27 항에 있어서,

상기 입력 표현은 텍스트, 멜-스케일(mel-scale) 스펙트로그램들, 기본 주파수(F0) 특징들, 또는 이들의 조합을 포함하는, 비일시적 컴퓨터 판독가능 저장 매체.

청구항 29

장치로서,

타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 수단; 및

상기 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 상기 하나 이상의 제어 파라미터들에 기초하여 상기 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 수단을 포함하는, 장치.

청구항 30

제 29 항에 있어서,

상기 수신하는 수단 및 상기 프로세싱하는 수단은 가상 어시스턴트, 홈 어플라이언스, 스마트 디바이스, 사물 인터넷 (IoT) 디바이스, 통신 디바이스, 헤드셋, 차량, 컴퓨터, 디스플레이 디바이스, 텔레비전, 게이밍 콘솔, 뮤직 플레이어, 라디오, 비디오 플레이어, 엔터테인먼트 유닛, 개인용 미디어 플레이어, 디지털 비디오 플레이어, 카메라, 또는 내비게이션 디바이스 중 적어도 하나에 통합되는, 장치.

발명의 설명

기술 분야

I. 우선권의 주장

본 출원은 2021 년 1 월 21 일 출원된 공동 소유된 미국 정규 특허출원 제 17/154,372 호로부터의 우선권의 이익을 주장하며, 이 출원의 내용들은 그 전부가 본 명세서에 참조로 통합된다.

II. 분야

본 개시는 일반적으로 합성된 스피치 생성에 관한 것이다.

배경 기술

기술의 발전은 더 더 작고 더 강력한 컴퓨팅 디바이스들을 유발하였다. 예를 들어, 모바일 및 스마트 폰들과 같은 무선 전화기들, 태블릿들 및 랩톱 컴퓨터들을 포함하는 다양한 휴대용 개인용 컴퓨팅 디바이스들이 현재 존재하고, 이들은 작고, 경량이며, 사용자들에 의해 용이하게 휴대된다. 이러한 디바이스들은 무선 네트워크들을 통해 음성 및 데이터 패킷들을 통신할 수 있다. 또한, 많은 이러한 디바이스들은 디지털 스틸 카메라, 디지털 비디오 카메라, 디지털 레코더, 및 오디오 파일 플레이어와 같은 부가적인 기능성을 통합한다.

또한, 이러한 디바이스들은 인터넷에 액세스하는데 사용될 수 있는, 웹 브라우저 애플리케이션과 같은 소프트웨어 애플리케이션들을 포함하는 실행가능한 명령들을 프로세싱할 수 있다. 이와 같이, 이들 디바이스들은 상당한 컴퓨팅 능력들을 포함할 수 있다.

이러한 컴퓨팅 디바이스들은 종종 하나 이상의 마이크로폰들로부터 오디오 신호를 수신하는 기능성을 통합한다.

예를 들어, 오디오 신호는 마이크로폰들에 의해 캡처된 사용자 스피치, 마이크로폰들에 의해 캡처된 외부 사운드들, 또는 이들의 조합을 나타낼 수도 있다. 이러한 디바이스들은 스피치 모델들을 사용하는 애플리케이션들을 포함할 수도 있다. 예를 들어, 음성-활성화 애플리케이션은 사용자 스피치 커맨드들을 인식한다.

스피치-기반 애플리케이션들은 오디오 스피치 샘플들을 사용하여 훈련된다. 훈련을 위해 오디오 스피치 샘플들의 더 크고 더 다양한 데이터 세트를 사용하는 것은 더 강건한 스피치 모델을 야기한다. 그러나, 사람으로부터 많은 수의 다양한 오디오 스피치 샘플들을 획득하는 것은 시간-소모적이고 불편할 수 있다.

발명의 내용

본 개시의 일 구현에 따르면, 스피치 생성을 위한 디바이스는 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하도록 구성된 하나 이상의 프로세서들을 포함한다. 하나 이상의 프로세서들은 또한, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하도록 구성된다.

본 개시의 다른 구현에 따르면, 스피치 생성 방법은, 디바이스에서, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 단계를 포함한다. 방법은 또한, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 단계를 포함한다.

본 개시의 다른 구현에 따르면, 비일시적 컴퓨터 판독가능 매체는, 하나 이상의 프로세서들에 의해 실행될 때, 하나 이상의 프로세서들로 하여금, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하게 하는 명령들을 저장한다. 명령들은, 하나 이상의 프로세서들에 의해 실행될 때, 또한 하나 이상의 프로세서들로 하여금, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프

로세싱하게 한다.

[0010] 본 개시의 다른 구현에 따르면, 장치는 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 수단을 포함한다. 장치는 또한, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 수단을 포함한다.

[0011] 본 개시의 다른 양태들, 이점들 및 특징들은 다음 섹션들: 도면의 간단한 설명, 상세한 설명 및 청구범위를 포함하여 출원 전체를 검토한 후 분명해질 것이다.

도면의 간단한 설명

[0012] 도 1 은 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 시스템의 특정한 예시적인 양태의 블록 다이어그램이다.

도 2 는 본 개시의 일부 예들에 따른, 도 1 의 시스템의 스피치 생성기의 특정한 구현의 다이어그램이다.

도 3 은 본 개시의 일부 예들에 따른, 도 1 의 스피치 생성기의 다른 특정한 구현의 다이어그램이다.

도 4a 는 본 개시의 일부 예들에 따른, 도 1 의 스피치 생성기의 다른 특정한 구현의 다이어그램이다.

도 4b 는 본 개시의 일부 예들에 따른, 도 1 의 스피치 생성기의 다른 특정한 구현의 다이어그램이다.

도 5 는 본 개시의 일부 예들에 따른, 도 1 의 스피치 생성기의 인코더의 예시적인 양태의 다이어그램이다.

도 6 은 본 개시의 일부 예들에 따른, 도 1 의 시스템의 컴포넌트들의 동작들의 예시적인 양태의 다이어그램이다.

도 7 은 본 개시의 일부 예들에 따른, 도 1 의 시스템의 컴포넌트들의 동작들의 예시적인 양태의 다이어그램이다.

도 8 은 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 집적 회로의 예를 도시한다.

도 9 는 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 모바일 디바이스의 다이어그램이다.

도 10 은 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 헤드셋의 다이어그램이다.

도 11 은 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 웨어러블 전자 디바이스의 다이어그램이다.

도 12 는 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 음성-제어형 스피커 시스템의 다이어그램이다.

도 13 은 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 카메라의 다이어그램이다.

도 14 는 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한, 가상 현실 또는 증강 현실 헤드셋과 같은 헤드셋의 다이어그램이다.

도 15 는 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 차량의 제 1 예의 다이어그램이다.

도 16 은 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 차량의 제 2 예의 다이어그램이다.

도 17 은 본 개시의 일부 예들에 따른, 도 1 의 디바이스에 의해 수행될 수도 있는 합성된 스피치 생성 방법의 특정한 구현의 다이어그램이다.

도 18 은 본 개시의 일부 예들에 따른, 도 1 의 디바이스에 의해 수행될 수도 있는 합성된 스피치 생성 방법의 다른 특정한 구현의 다이어그램이다.

도 19 는 본 개시의 일부 예들에 따른, 합성된 스피치를 생성하도록 동작가능한 디바이스의 특정한 예시적인 양태의 블록 다이어그램이다.

발명을 실시하기 위한 구체적인 내용

- [0013] 훈련을 위해 오디오 스피치 샘플들의 더 크고 더 다양한 데이터 세트를 사용하는 것은 더 강건한 스피치 모델들을 야기한다. 스피치 훈련을 위해 많은 수의 오디오 스피치 샘플들을 제공하는 것은 사용자들에게 불편하고 시간 소모적일 수 있다. 합성 스피치 생성 시스템들 및 방법들이 개시된다. 예를 들어, 스피치 생성기는 스피치의 입력 표현 (예를 들어, 오디오 신호, 비디오 신호, 텍스트 등) 을 수신 (또는 액세스) 한다. 스피치 생성기는 타겟 감정 (예를 들어, 중립, 행복, 슬픔 등), 타겟 스피치 레이트 (예를 들어, 느림, 중간, 또는 빠름), 스피치 특성들이 사용될 타겟 사람 (예를 들어, 사용자, 유명인, 특정한 억양을 갖는 어린이 등), 또는 이들의 조합과 같은 타겟 스피치 특성들을 표시하는 제어 파라미터들을 수신한다. 스피치 생성기는 타겟 스피치 특성들에 기초하여 인코딩된 데이터를 생성하기 위해 제어 파라미터들에 기초하여 입력 표현을 프로세싱하는 멀티-인코더 변환기와 같은 멀티-인코더, 및 인코딩된 데이터를 디코딩함으로써 오디오 신호 (예를 들어, 합성 스피치 신호) 를 생성하는 디코더를 포함한다. 스피치 생성기는 제어 파라미터들의 다양한 세트들을 동일한 입력 표현에 적용함으로써 합성 오디오 샘플들의 다양한 세트를 생성할 수 있다.
- [0014] 본 개시의 특정한 양태들은 도면들을 참조하여 하기에서 설명된다. 설명에서, 공통 특징들은 공통 참조 번호들로 지정된다. 본 명세서에서 사용된 바와 같이, 다양한 용어가 단지 특정한 구현들을 설명하는 목적으로 사용되며 구현들을 한정하는 것으로 의도되지 않는다. 예를 들어, 단수 형태들 "a", "an" 및 "the" 는, 문맥이 달리 분명히 나타내지 않는 한, 복수 형태들을 물론 포함하도록 의도된다. 또한, 본 명세서에 설명된 일부 특징들은 일부 구현들에서 단수이고 다른 구현들에서는 복수이다. 예시를 위해, 도 1 은 하나 이상의 프로세서들 (도 1 의 "프로세서(들)"(190)) 을 포함하는 디바이스 (102) 를 도시하며, 이는 일부 구현들에서, 디바이스 (102) 가 단일 프로세서 (190) 를 포함하고, 다른 구현들에서, 디바이스 (102) 가 다중 프로세서들 (190) 을 포함함을 나타낸다. 본 명세서에서의 참조를 용이하게 하기 위해, 이러한 특징들은 일반적으로 "하나 이상의" 특징으로서 도입되고, 후속하여 특징들 중 다중 특징들과 관련된 양태들이 설명되지 않는 한 단수로 지칭된다.
- [0015] 본 명세서에서 사용된 바와 같이, 용어들 "포함한다(comprise)", "포함한다(comprises)", 및 "포함하는(comprising)" 은 "포함한다(include)", "포함한다(includes)", 또는 "포함하는(including)" 과 상호교환가능하게 사용될 수도 있다. 부가적으로, 용어 "여기서 (wherein)" 는 "여기에서 (where)" 와 상호교환가능하게 사용될 수도 있다. 본 명세서에서 사용된 바와 같이, "예시적인" 은 예, 구현, 및/또는 양태를 나타내며, 제한하는 것으로서 또는 선호도 또는 선호된 구현을 나타내는 것으로서 해석되지 않아야 한다. 본 명세서에 사용된 바와 같이, 구조, 컴포넌트, 동작 등과 같은 엘리먼트를 수정하는데 사용되는 서수 용어 (예를 들어, "제 1", "제 2", "제 3" 등) 는 그것만으로도 그 엘리먼트의 다른 엘리먼트에 대한 어떤 우선순위 또는 순서도 표시하지 않고, 오히려 그 엘리먼트를 (서수 용어의 사용이 없다면) 동일 명칭을 갖는 다른 엘리먼트와 구별할 뿐이다. 본 명세서에서 사용된 바와 같이, 용어 "세트 (set)" 는 특정한 엘리먼트의 하나 이상을 지칭하고, 용어 "복수" 는 특정한 엘리먼트의 배수 (예를 들어, 2개 이상) 를 지칭한다.
- [0016] 본 명세서에서 사용된 바와 같이, "커플링된" 은 "통신가능하게 커플링된", "전기적으로 커플링된", 또는 "물리적으로 커플링된" 을 포함할 수도 있고, 또한 (또는 대안으로) 이들의 임의의 조합을 포함할 수도 있다. 2개의 디바이스들 (또는 컴포넌트들) 은 하나 이상의 다른 디바이스, 컴포넌트, 와이어, 버스, 네트워크 (예를 들어, 유선 네트워크, 무선 네트워크, 또는 이들의 조합) 등을 통해 직접적으로 또는 간접적으로 커플링 (예를 들어, 통신가능하게 커플링, 전기적으로 커플링, 또는 물리적으로 커플링) 될 수도 있다. 전기적으로 커플링된 2개의 디바이스들 (또는 컴포넌트들) 은 동일한 디바이스 또는 상이한 디바이스들에 포함될 수도 있고, 예시적인, 비-제한적인 예들로서 전자기기들, 하나 이상의 커넥터 또는 유도 커플링을 통해 연결될 수도 있다. 일부 구현들에서, 전기 통신에서와 같이, 통신가능하게 커플링되는 2개의 디바이스들 (또는 컴포넌트들) 은 하나 이상의 와이어, 버스, 네트워크 등을 통해, 직접 또는 간접적으로 신호들 (예를 들어, 디지털 신호들 또는 아날로그 신호들) 을 전송 및 수신할 수도 있다. 본 명세서에서 사용된 바와 같이, "직접 커플링된" 은 컴포넌트들을 개재하지 않으면서 커플링 (예를 들어, 통신가능하게 커플링, 전기적으로 커플링, 또는 물리적으로 커플링) 되는 2개의 디바이스들을 포함할 수도 있다.
- [0017] 본 개시에서, "결정하는 것", "계산하는 것", "추정하는 것", "시프트하는 것", "조정하는 것" 등과 같은 용어들은 하나 이상의 동작이 어떻게 수행되는지를 설명하기 위해 사용될 수도 있다. 이러한 용어들은 한정하는 것으로서 해석되지 않아야 하고 다른 기법들이 유사한 동작들을 수행하는데 활용될 수도 있음에 유의해야 한다. 부가적으로, 본 명세서에서 언급된 바와 같이, "생성하는 것", "계산하는 것", "추정하는 것", "사용하는

것", "선택하는 것", "액세스하는 것", 및 "결정하는 것"은 상호교환가능하게 사용될 수도 있다. 예를 들어, 파라미터 (또는 신호)를 "생성하는 것", "계산하는 것", "추정하는 것" 또는 "결정하는 것"은 파라미터 (또는 신호)를 능동으로 생성하는 것, 추정하는 것, 계산하는 것, 또는 결정하는 것을 지칭할 수도 있거나 또는 다른 컴포넌트 또는 디바이스에 의해서와 같이, 이미 생성된 파라미터 (또는 신호)를 사용하는 것, 선택하는 것, 또는 액세스하는 것을 지칭할 수도 있다.

[0018] 도 1을 참조하면, 합성된 스피치를 생성하도록 구성된 시스템의 특정한 예시적인 양태가 개시되고 일반적으로 100으로 지정된다. 시스템 (100)은 스피치 생성기 (120)를 사용하여 합성 스피치를 생성하도록 구성된 디바이스 (102)를 포함한다. 디바이스 (102)는 메모리 (132)에 커플링된 하나 이상의 프로세서들 (190)을 포함한다. 하나 이상의 프로세서들 (190)은 스피치 생성기 (120)를 포함한다. 일부 구현들에서, 디바이스 (102)는 마이크로폰 (182), 스피치 저장소 (184), 또는 양자 모두에 커플링된다. 특정한 구현에서, 마이크로폰 (182), 스피치 저장소 (184), 또는 양자 모두는 디바이스 (102) 외부에 있다. 대안의 구현에서, 마이크로폰 (182), 스피치 저장소 (184), 또는 양자 모두는 디바이스 (102)에 통합된다.

[0019] 스피치 생성기 (120)는 오디오 분석기 (122), 스타일 데이터 생성기 (126), 멀티-인코더 변환기 (128), 스피치 합성기 (130), 또는 이들의 조합을 포함한다. 특정한 양태에서, 오디오 분석기 (122)는 스타일 데이터 생성기 (126), 멀티-인코더 변환기 (128), 또는 양자 모두에 커플링된다. 스타일 데이터 생성기 (126)는 멀티-인코더 변환기 (128)를 통해 스피치 합성기 (130)에 커플링된다. 특정한 양태에서, 스피치 합성기 (130)는 스피치 저장소 (184)에 커플링된다.

[0020] 오디오 분석기 (122)는 입력 신호 (121)를 프로세싱하여 스피치의 입력 표현 (123)을 생성하도록 구성된다. 특정한 예에서, 입력 신호 (121)는 예를 들어, 마이크로폰 (182)을 통해 수신되거나 저장 디바이스로부터 취출된 오디오 신호를 포함한다. 다른 예에서, 입력 신호 (121)는 오디오 정보 및 이미지 정보를 나타내는 비디오 신호를 포함한다. 입력 표현 (123)은 입력 신호 (121)에 기초하여 텍스트, 멜-스케일 스펙트로그램들, 기본 주파수 (F0) 특징들, 스피치의 하나 이상의 다른 표현들, 또는 이들의 조합을 포함한다. 특정한 예에서, 오디오 분석기 (122)는 입력 표현 (123)으로서, 멜-스케일 스펙트로그램들, 기본 주파수 (F0) 특징들, 또는 이들의 조합을 생성하기 위해 특징 추출을 수행하도록 구성된다. 특정한 예에서, 오디오 분석기 (122)는 입력 표현 (123)으로서 텍스트를 생성하기 위해 입력 신호 (121)에 대해 스피치-텍스트 변환을 수행하도록 구성된다.

[0021] 스타일 데이터 생성기 (126)는, 도 2를 참조하여 추가로 설명된 바와 같이, 입력 표현 (123) 및 하나 이상의 제어 파라미터들 (125)에 기초하여 병합된 스타일 데이터 (127)를 생성하도록 구성된다. 하나 이상의 제어 파라미터들 (125)은 하나 이상의 타겟 스피치 특성들 (141)을 나타낸다. 예를 들어, 타겟 스피치 특성들 (141)은 타겟 사람 (143)(예를 들어, 사람 (180)), 타겟 감정 (145)(예를 들어, 행복, 분노, 흥분, 슬픔, 또는 중립 감정), 타겟 스피치 레이트 (147)(예를 들어, 스피치의 중간 레이트), 또는 이들의 조합을 표시한다. 특정한 양태에서, 병합된 스타일 데이터 (127)는 입력 표현 (123) 및 하나 이상의 제어 파라미터들 (125)을 표현하는 임베딩 (예를 들어, 임베딩 벡터)에 대응한다.

[0022] 멀티-인코더 변환기 (128)는 도 4a 내지 도 5를 참조하여 추가로 설명된 바와 같이, 출력 스펙트럼 데이터 (129)를 생성하기 위해 입력 표현 (123) 및 병합된 스타일 데이터 (127)를 프로세싱하도록 구성된다. 멀티-인코더 변환기 (128)는 멀티-인코더 (192) 및 디코더 (194)를 포함한다. 멀티-인코더 (192)는 합성된 스피치 신호 (131)에 대응하는 인코딩된 데이터 (193)를 생성하기 위해 하나 이상의 제어 파라미터들 (125)에 기초하여 입력 표현 (123)을 프로세싱하도록 구성된다. 예를 들어, 합성된 스피치 신호 (131)는 인코딩된 데이터 (193)에 기초한다. 특정한 양태에서, 멀티-인코더 (192)는 도 4a를 참조하여 추가로 설명된 바와 같이, 제 1 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들 (125)과 독립적으로 (예를 들어, 타겟 스피치 특성들 (141)과 독립적으로) 입력 표현 (123)을 프로세싱하도록 구성된 제 1 인코더를 포함한다. 멀티-인코더 (192)는 도 4a 및 도 4b를 참조하여 추가로 설명된 바와 같이, 제 2 인코딩된 데이터를 생성하기 위해 (하나 이상의 제어 파라미터들 (125)에 기초하는) 병합된 스타일 데이터 (127)를 프로세싱하도록 구성된 하나 이상의 제 2 인코더들을 포함한다. 따라서, 제 2 인코딩된 데이터는 하나 이상의 제어 파라미터들 (125)(예를 들어, 타겟 스피치 특성들(141))에 기초한다. 디코더 (194)는 도 4a 및 도 4b를 참조하여 추가로 설명된 바와 같이, 인코딩된 데이터 (193)(예를 들어, 제 1 인코딩된 데이터 및 제 2 인코딩된 데이터)에 기초하여 출력 스펙트럼 데이터 (129)를 생성하도록 구성된다. 스피치 합성기 (130)는 타겟 스피치 특성들 (141)에 기초하여 (예를 들어, 입력 표현 (123)에 의해 표시된) 스피치의 버전을 표현하는 합성된 스피치 신호 (131)를 생성하기 위해 출력 스펙트럼 데이터 (129)에 대해 스피치 합성을 수행하도록 구성

된다.

- [0023] 메모리 (132) 는 스피치 생성기 (120) 에 의해 사용되거나 생성된 데이터를 저장하도록 구성된다. 예를 들어, 메모리 (132) 는 입력 신호 (121), 입력 표현 (123), 하나 이상의 제어 파라미터들 (125), 병합된 스타일 데이터 (127), 인코딩된 데이터 (193), 출력 스펙트럼 데이터 (129), 합성된 스피치 신호 (131) 의 적어도 일부, 또는 이들의 조합의 적어도 일부를 저장하도록 구성된다. 특정한 양태에서, 메모리 (132) 는 하나 이상의 제어 파라미터들 (125) 과 타겟 스피치 특성들 (141) 사이의 매핑을 표시하는 매핑 데이터를 저장한다. 예를 들어, 매핑 데이터는 하나 이상의 제어 파라미터들 (125) 의 제 1 제어 파라미터의 값들과 사람들의 세트 (151) 사이의 제 1 매핑들, 하나 이상의 제어 파라미터들 (125) 의 제 2 제어 파라미터의 값들과 감정들의 세트 (161) 사이의 제 2 매핑들, 하나 이상의 제어 파라미터들 (125) 의 제 3 제어 파라미터의 값들과 스피치의 레이트들의 세트 (171) 사이의 제 3 매핑들, 또는 이들의 조합을 표시한다. 스타일 데이터 생성기 (126) 는 매핑 데이터가 하나 이상의 제어 파라미터들 (125) 에 의해 표시된 제 1 제어 파라미터의 값이 사람들의 세트 (151) 의 타겟 사람 (143) 에 매핑됨을 표시한다고 결정하는 것에 응답하여, 하나 이상의 제어 파라미터들 (125) 이 타겟 사람 (143) 을 표시한다고 결정한다. 유사하게, 스타일 데이터 생성기 (126) 는 매핑 데이터가 하나 이상의 제어 파라미터들 (125) 에 의해 표시된 제 2 제어 파라미터의 값이 감정들의 세트 (161) 의 타겟 감정 (145) 에 매핑됨을 표시한다고 결정하는 것에 응답하여 하나 이상의 제어 파라미터들 (125) 이 타겟 감정 (145) 을 표시한다고 결정한다. 스타일 데이터 생성기 (126) 는 매핑 데이터가 하나 이상의 제어 파라미터들 (125) 에 의해 표시된 제 3 제어 파라미터의 값이 스피치의 레이트들의 세트 (171) 중 타겟 스피치 레이트 (147) 에 매핑됨을 표시한다고 결정하는 것에 응답하여 하나 이상의 제어 파라미터들 (125) 이 타겟 스피치 레이트 (147) 를 표시한다고 결정한다.
- [0024] 일부 구현들에서, 디바이스 (102) 는 하나 또는 다양한 타입들의 디바이스들에 대응하거나 그에 포함된다. 예시적인 예에서, 하나 이상의 프로세서들 (190) 은 도 10 을 참조하여 추가로 설명된 바와 같이, 마이크로폰 (182) 을 포함하는 헤드셋 디바이스에 통합된다. 다른 예들에서, 하나 이상의 프로세서들 (190) 은 도 9 를 참조하여 설명된 바와 같은 모바일 폰 또는 태블릿 컴퓨터 디바이스, 도 11 을 참조하여 설명된 바와 같은 웨어러블 전자 디바이스, 도 12 를 참조하여 설명된 바와 같은 음성-제어형 스피커 시스템, 도 13 을 참조하여 설명된 바와 같은 카메라 디바이스, 또는 도 14 를 참조하여 설명된 바와 같은 가상 현실, 증강 현실, 또는 혼합 현실 헤드셋 중 적어도 하나에 통합된다. 다른 예시적인 예에서, 하나 이상의 프로세서 (190) 는 도 15 및 도 16 을 참조하여 추가로 설명된 바와 같은, 마이크로폰 (182) 을 또한 포함하는 차량에 통합된다.
- [0025] 동작 동안, 오디오 분석기 (122) 는 마이크로폰 (182) 을 통해 입력 신호 (121) 를 수신한다. 예를 들어, 마이크로폰 (182) 은 사람 (180) 의 스피치 (예를 들어, "도어 잠금해제") 를 캡처하고, 스피치를 표현하는 입력 신호 (121) 를 오디오 분석기 (122) 에 제공한다. 입력 신호 (121) 에 의해 표현된 스피치 (예를 들어, "도어 잠금해제") 는 특정한 스피치의 레이트 (예를 들어, 스피치의 빠른 레이트) 및 특정한 감정 (예를 들어, 불안함) 을 갖는다.
- [0026] 오디오 분석기 (122) 는 입력 신호 (121) 의 입력 표현 (123) 을 생성한다. 예를 들어, 오디오 분석기 (122) 는 입력 신호 (121) 에 대해 스피치-투-텍스트 변환을 수행하여 텍스트 (예를 들어, "도어 잠금해제") 를 생성한다. 다른 예로서, 오디오 분석기 (122) 는 멜-스케일 스펙트로그램들, 기본 주파수 (F0) 특징들, 또는 이들의 조합을 생성하기 위해 입력 신호 (121) 에 대해 오디오 특징 추출을 수행한다. 입력 표현 (123) 은 텍스트 (예를 들어, "도어 잠금해제"), 멜-스케일 스펙트로그램들, 기본 주파수 (F0) 특징들 또는 이들의 조합을 포함한다.
- [0027] 디바이스 (102) 는 하나 이상의 제어 파라미터들 (125) 을 수신하거나 액세스한다. 특정한 양태에서, 하나 이상의 제어 파라미터들 (125) 은 디폴트 데이터, 구성 설정, 사용자 입력, 또는 이들의 조합에 기초한다. 하나 이상의 제어 파라미터들 (125) 은 스피치 생성기 (120) 에 의해 생성될 합성된 스피치의 타겟 스피치 특성들 (141) 을 표시한다.
- [0028] 특정한 양태에서, 타겟 스피치 특성들 (141) 은 사람들의 세트 (151) 의 타겟 사람 (143)(예를 들어, 사람 (180)) 을 표시한다. 사람들의 세트 (151) 는 제 1 사람 (사람 1)(153), 제 2 사람(사람 2)(155), 제 N 사람 (사람 N)(157) 과 같은, 하나 이상의 부가 사람들, 또는 이들의 조합을 포함한다. 예시적이고 비제한적인 예에서, 사람들의 세트 (151) 는 특정한 유명인, 특정한 사용자, 특정한 역할을 갖는 사람, 특정한 연령의 사람, 특정한 성별의 사람, 또는 이들의 조합을 포함한다. 타겟 사람 (143) 의 스피치 특성들은 합성된 스피치를 생성하기 위해 스피치 생성기 (120) 에 의해 사용될 것이다. 특정한 예에서, 타겟 사람 (143) 은 "

오바마 대통령” 을 표시한다. 다른 예에서, 타겟 사람 (143) 은 "텍사스 억양을 갖는 중년 남성" 을 표시한다. 특정한 예에서, 타겟 사람 (143) 은 사람 (180) 을 표시한다.

[0029] 특정한 양태에서, 타겟 스피치 특성들 (141) 은 감정들의 세트 (161) 로부터 선택된 타겟 감정 (145)(예를 들어, 중립) 을 표시한다. 감정들의 세트 (161) 는 제 1 감정 (감정 1)(163), 제 2 감정 (감정 2)(165), 제 N 감정 (감정 N)(167) 과 같은, 하나 이상의 부가 사람들, 또는 이들의 조합을 포함한다. 예시적이고 비제한적인 예에서, 감정들의 세트 (161) 는 행복, 공포, 불안, 흥분, 분노, 평온, 중립, 또는 이들의 조합을 포함한다.

[0030] 특정한 양태에서, 타겟 스피치 특성들 (141) 은 스피치의 레이트들의 세트 (171) 로부터 선택된 타겟 스피치 레이트 (147)(예를 들어, 스피치의 중간 레이트) 를 표시한다. 스피치의 레이트들의 세트 (171) 는 스피치의 제 1 레이트 (스피치의 레이트 1)(173), 스피치의 제 2 레이트 (스피치의 레이트 2)(175), 스피치의 제 N 레이트 (스피치의 레이트 N)(177) 와 같은 스피치의 하나 이상의 부가적인 레이트들, 또는 이들의 조합을 포함한다. 예시적이고 비제한적인 예에서, 스피치의 레이트들의 세트 (171) 는 스피치의 느린 레이트, 스피치의 중간 레이트, 스피치의 빠른 레이트, 또는 이들의 조합을 포함한다.

[0031] 스타일 데이터 생성기 (126) 는, 도 2 를 참조하여 추가로 설명된 바와 같이, 입력 표현 (123) 및 하나 이상의 제어 파라미터들 (125) 에 기초하여 병합된 스타일 데이터 (127) 를 생성한다. 예를 들어, 병합된 스타일 데이터 (127) 는 입력 표현 (123) 및 하나 이상의 제어 파라미터들 (125) 을 표현하는 임베딩 (예를 들어, 임베딩 벡터) 에 대응한다.

[0032] 멀티-인코더 변환기 (128) 는 도 4a 및 도 4b 를 참조하여 추가로 설명된 바와 같이, 출력 스펙트럼 데이터 (129) 를 생성하기 위해 입력 표현 (123) 및 병합된 스타일 데이터 (127) 를 프로세싱한다. 예를 들어, 멀티-인코더 (192) 의 제 1 인코더는 도 4a 를 참조하여 추가로 설명된 바와 같이, 제 1 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들 (125) 과 독립적으로 (예를 들어, 타겟 스피치 특성들 (141) 과 독립적으로) 입력 표현 (123) 을 프로세싱한다. 멀티-인코더 (192) 의 하나 이상의 제 2 인코더들은 도 4a 및 도 4b 를 참조하여 추가로 설명된 바와 같이, 제 2 인코딩된 데이터를 생성하기 위해 병합된 스타일 데이터 (127) 를 프로세싱한다.

[0033] 특정한 양태에서, 병합된 스타일 데이터 (127) 는 입력 표현 (123) 의 프로세싱 동안 멀티-인코더 (192) 에 의해 사용된다. 예를 들어, 하나 이상의 제 2 인코더들은 입력 표현 (123) 을 프로세싱하는 제 1 인코더와 동시에 병합된 스타일 데이터 (127) 를 프로세싱한다. 멀티-인코더 (192) 는 인코딩된 데이터 (193) (예를 들어, 제 1 인코딩된 데이터, 제 2 인코딩된 데이터, 또는 이들의 조합) 를 디코더 (194) 에 제공한다. 특정한 양태에서, 인코딩된 데이터 (193) 는 멀티-인코더 (192) 의 출력 임베딩 (예를 들어, 임베딩 벡터) 에 대응한다.

[0034] 디코더 (194) 는 도 4a 를 참조하여 추가로 설명된 바와 같이, 출력 스펙트럼 데이터 (129) 를 생성하기 위해 인코딩된 데이터 (193)(예를 들어, 제 1 인코딩된 데이터 및 제 2 인코딩된 데이터) 를 프로세싱한다. 스피치 합성기 (130) 는 합성된 스피치 신호 (131) 를 생성하기 위해 출력 스펙트럼 데이터 (129) 에 대해 스피치 합성을 수행한다. 합성된 스피치 신호 (131) 는 타겟 스피치 특성들 (141) 에 기초하는 입력 표현 (123) 에 의해 표시된 스피치의 버전에 대응한다.

[0035] 멀티-인코더 (192) 및 디코더 (194)는 예시적인 예로서 멀티-인코더 변환기 (128) 에 포함되는 것으로 설명된다. 일부 예들에서, 멀티-인코더 (192) 및 디코더 (194) 는 변환기 이외의 구성을 갖는다.

[0036] 하나 이상의 제어 파라미터들 (125) 에 대해 (그리고 하나 이상의 제어 파라미터들 (125) 과 독립적인 제 1 인코더와는 별도로) 멀티-인코더 (192) 에 하나 이상의 제 2 인코더들을 갖는 것은, 도 3 을 참조하여 추가로 설명된 바와 같이, 제 2 인코더들이 하나 이상의 제어 파라미터들 (125) 의 각각에 기초하여 입력 표현들 (123) 을 인코딩하기 위해 특별히 훈련 (예를 들어, 최적화) 되는 것을 가능하게 한다.

[0037] 특정한 예에서, 합성된 스피치 신호 (131) 는 타겟 스피치 특성들 (141) 을 갖는다. 특정한 예에서, 합성된 스피치 신호 (131) 는 (입력 신호 (121) 와 비교하여) 타겟 스피치 특성들 (141) 을 갖는 것에 더 가깝다. 특정한 양태에서, 스피치 생성기 (120) 는 도 3 을 참조하여 추가로 설명되는 바와 같이, 합성된 스피치 신호 (131) 에 대응하는 추정된 제어 파라미터들과 하나 이상의 제어 파라미터들 (125) 의 비교에 기초하여 멀티-인코더 변환기 (128) 의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 을 조정하도록 구성되는 조정기를 포함한다. 따라서, 스피치 생성기 (120) 는 합성 스피치 생성의 정확도를 개선하기 위해 멀티-인코더 변환

기 (128) 를 동적으로 훈련할 수 있다.

[0038] 디바이스 (102) 는 예시적인 예로서 입력 신호 (121) 를 수신하고 입력 신호 (121) 에 기초하여 입력 표현 (123) 을 생성하는 것으로 설명된다. 예를 들어, 스타일 데이터 생성기 (126) 및 멀티-인코더 변환기 (128) 의 각각은 오디오 분석기 (122) 로부터 입력 표현 (123) 을 수신한다. 일부 예들에서, 디바이스 (102) 는 (예를 들어, 입력 신호 (121) 대신에) 입력 표현 (123) 을 수신한다. 예를 들어, 디바이스 (102)(예를 들어, 멀티-인코더 변환기 (128) 및 스타일 데이터 생성기 (126) 각각) 는 사람 (180) 으로부터, 다른 디바이스 로부터 또는 양자 모두로부터 입력으로서 입력 표현 (123) 을 수신한다.

[0039] 예시적인 예에서, 입력 표현 (123) 은 특정한 감정 (예를 들어, 불안함) 및 특정한 스피치의 레이트 (예를 들어, 스피치의 높은 레이트) 를 갖는 사람 (180) 의 스피치에 대응한다. 합성된 스피치 신호 (131) 는 타겟 감정 (145)(예를 들어, 중립) 및 타겟 스피치 레이트 (147)(예를 들어, 스피치의 중간 레이트) 를 갖는 타겟 사람 (143)(예를 들어, 사람 (180)) 의 스피치에 대응한다. 이 예에서, 합성된 스피치 신호 (131) 에 대해 스피치 인식 (예를 들어, 스피치의 중간 레이트를 갖는 중립 스피치) 을 수행하는 것은 입력 신호 (121) 에 대해 스피치 인식 (예를 들어, 스피치의 높은 레이트를 갖는 불안한 스피치) 을 수행하는 것에 비해 인식 정확도를 개선한다.

[0040] 특정한 양태에서, 스피치 생성기 (120) 는 합성된 스피치 신호 (131) 를 스피치 저장소 (184) 에 부가한다. 예를 들어, 스피치 생성기 (120) 는 도 6 을 참조하여 추가로 설명된 바와 같이, 입력 신호 (121) 에 기초하여 훈련 데이터셋을 생성 (또는 업데이트) 한다. 예시를 위해, 스피치 생성기 (120) 는 하나 이상의 제어 파라미터들 (125) 의 제 1 세트에 기초하여 제 1 합성된 스피치 신호 (131), 하나 이상의 제어 파라미터들 (125) 의 제 2 세트에 기초하여 제 2 합성된 스피치 신호 (131), 하나 이상의 제어 파라미터들 (125) 의 하나 이상의 부가 세트들에 기초하여 하나 이상의 부가 합성된 스피치 신호들 (131), 또는 이들의 조합을 생성한다. 스피치 생성기 (120) 는 입력 신호 (121), 제 1 합성된 스피치 신호 (131), 제 2 합성된 스피치 신호 (131), 하나 이상의 부가 합성된 스피치 신호들 (131), 또는 이들의 조합을 스피치 저장소 (184) 에서의 훈련 데이터셋에 부가한다.

[0041] 특정한 예에서, 스피치 생성기 (120) 는 동일한 입력 신호 (121) 에 기초하여 스피치의 다양한 감정들 및 레이트들에 대응하는 다중의 합성된 스피치 신호들을 생성한다. 이 예에서, 합성된 스피치 신호들은 사람 (180) 에 의해 스피킹된 몇몇 단어들을 사용하여 사람 (180) 의 스피치를 표현하는 강건한 스피치 모델 (예를 들어, 뉴럴 네트워크) 을 훈련하는데 사용될 수 있다. 예시를 위해, 등록 동안, 특정한 스피치 특성들 (예를 들어, 스피치의 중간 레이트를 갖는 중립 스피치) 을 갖는 사람 (180) 의 스피치 (예를 들어, 등록 스피치) 가 뉴럴 네트워크를 훈련하기하기 위한 입력 신호 (121) 로서 수신된다. 다양한 스피치 특성들에 대응하는 합성된 스피치에 기초하여 뉴럴 네트워크를 훈련시키면, 뉴럴 네트워크가 등록 스피치의 스피치 특성들 (예를 들어, 스피치의 중간 레이트를 갖는 중립 스피치) 과는 상이한 스피치 특성들 (예를 들어, 스피치의 빠른 레이트를 갖는 불안한 스피치) 을 갖는 스피치를 인식하는 것을 가능하게 한다.

[0042] 특정한 예에서, 스피치 생성기 (120) 는 동일한 입력 신호 (121) 에 기초하여 다양한 사람들의 스피치 특성들, 스피치의 다양한 레이트들 및 다양한 감정들에 대응하는 다중의 합성된 스피치 신호들을 생성한다. 이 예에서, 크고 다양한 합성된 스피치 데이터셋은 사람 (180) 에 의해 스피킹된 몇몇 단어들을 사용하여 생성될 수 있다.

[0043] 마이크로폰 (182) 이 디바이스 (102) 에 커플링되는 것으로서 예시되지만, 다른 구현들에서는 마이크로폰 (182) 이 디바이스 (102) 에 통합될 수도 있다. 하나의 마이크로폰 (182) 이 예시되어 있지만, 다른 구현들에서는, 사용자 스피치를 캡처하도록 구성된 하나 이상의 부가 마이크로폰들이 포함될 수도 있다. 일부 구현들에서, 예컨대 스피치 생성기 (120) 가 (예를 들어, 메모리 (132) 로부터 추출되거나 다른 디바이스로부터 수신된 텍스트 데이터로서) 스토리지로부터 입력 신호 (121) 또는 입력 표현 (123) 을 추출할 때, 마이크로폰 (182) 은 생략된다. 시스템 (100) 이 디바이스 (102) 에서 스피치 생성기 (120) 의 모든 컴포넌트들을 포함하는 것으로 예시되어 있지만, 다른 구현들에서는, 제 2 디바이스가 부가될 수도 있고, 제 2 디바이스는 디바이스 (102) 에서 수행되는 것으로 설명된 동작들 중 일부를 수행할 수도 있다.

[0044] 도 2 를 참조하면, 스피치 생성기 (120) 의 특정한 구현이 나타나 있다. 스타일 데이터 생성기 (126) 는 사람 인코더 (206), 감정 인코더 (208), 스피치 레이트 인코더 (210), 하나 이상의 부가 스피치 특성 인코더들, 또는 이들의 조합과 같은 복수의 스피치 특성 인코더들을 포함한다.

- [0045] 특정한 양태에서, 오디오 분석기 (122) 는 복수의 스피치 특성 인코더들의 각각에 커플링된다. 복수의 스피치 특성 인코더들의 각각은 인코더 프리-네트 (212) 에 커플링된다. 복수의 스피치 특성 인코더들의 각각은 하나 이상의 제어 파라미터들 (125) 에 의해 표시된 대응하는 스피치 특성에 기초하여 입력 표현 (123) 의 인코딩된 버전을 생성하도록 구성된다. 예를 들어, 사람 인코더 (206) 는 사람 인코딩된 입력 표현 (207) 을 생성하기 위해 입력 표현 (123) 및 하나 이상의 제어 파라미터들 (125) 의 사람 제어 파라미터 (예를 들어, 타겟 사람 (143) 을 표시함) 를 프로세싱하도록 구성된다. 예시를 위해, 사람 인코더 (206) 는 입력 표현 (123) 및 사람 제어 파라미터 (예를 들어, 타겟 사람 (143) 을 표시함) 를 입력들로서 수신하고 입력 표현 (123) 및 사람 제어 파라미터를 표현하는 사람 인코딩된 입력 표현 (207)(예를 들어, 임베딩 벡터) 을 출력하도록 구성된 뉴럴 네트워크 (예를 들어, 컨볼루셔널 뉴럴 네트워크 (CNN)) 를 포함한다.
- [0046] 유사하게, 감정 인코더 (208) 는 감정 인코딩된 입력 표현 (209) 을 생성하기 위해 하나 이상의 제어 파라미터들 (125) 의 감정 제어 파라미터 (예를 들어, 타겟 감정 (145) 을 표시함) 및 입력 표현 (123) 을 프로세싱하도록 구성된다. 특정한 예에서, 스피치 레이트 인코더 (210) 는 스피치 레이트 인코딩된 입력 표현 (211) 을 생성하기 위해 하나 이상의 제어 파라미터들 (125) 의 스피치 파라미터의 레이트 (예를 들어, 타겟 스피치 레이트 (147) 를 표시함) 및 입력 표현 (123) 을 프로세싱하도록 구성된다.
- [0047] 인코더 프리-네트 (212) 는 뉴럴 네트워크 (예를 들어, CNN), 선형 프로젝션 계층, 또는 양자 모두를 포함한다. 특정한 예에서, 사람 인코딩된 입력 표현 (207), 감정 인코딩된 입력 표현 (209), 스피치 레이트 인코딩된 입력 표현 (211), 또는 이들의 조합은 뉴럴 네트워크 (예를 들어, CNN) 의 입력 임베딩에 대응한다. 인코더 프리-네트 (212) 의 뉴럴 네트워크는 출력 임베딩을 생성하기 위해 입력 임베딩 (예를 들어, 인코딩된 입력 표현) 을 프로세싱한다. 뉴럴 네트워크는 출력 임베딩을 선형 프로젝션 계층에 제공하고 선형 프로젝션 계층은 뉴럴 네트워크의 출력 임베딩에 대해 선형 프로젝션을 수행함으로써 병합된 스타일 데이터 (127) 를 생성한다. 특정한 양태에서, 멀티-인코더 변환기 (128) 는 어떠한 컨볼루션도 포함하지 않으며, 사람 인코딩된 입력 표현 (207), 감정 인코딩된 입력 표현 (209), 스피치 레이트 인코딩된 입력 표현 (211), 또는 이들의 조합을 프로세싱하기 위해 인코더 프리-네트 (212) 의 사용을 통해 컨볼루션이 수행된다. 인코더 프리-네트 (212) 는 병합된 스타일 데이터 (127) 를 멀티-인코더 (192) 에 제공한다.
- [0048] 멀티-인코더 (192) 는 오디오 분석기 (122) 로부터 입력 표현 (123) 을 수신하고 스타일 데이터 생성기 (126) 로부터 병합된 스타일 데이터 (127) 를 수신한다. 멀티-인코더 (192) 는 입력 표현 (123) 및 병합된 스타일 데이터 (127) 를 프로세싱하여 인코딩된 데이터 (193) 를 생성한다. 예를 들어, 멀티-인코더 (192) 는 도 4a 및 도 4b 를 참조하여 추가로 설명된 바와 같이, 입력 표현 (123) 을 프로세싱하여 도 1 의 타겟 스피치 특성들 (141) 과는 독립적인 제 1 인코딩된 데이터를 생성하고, 병합된 스타일 데이터 (127) 를 프로세싱하여 도 1 의 타겟 스피치 특성들 (141) 에 대응하는 제 2 인코딩된 데이터를 생성한다. 특정한 양태에서, 인코더 프리-네트 (212) 는 또한 병합된 스타일 데이터 (127) 를 디코더 (194) 에 제공한다. 이러한 양태에서, 디코더 (194) 는 출력 스펙트럼 데이터 (129) 를 생성하기 위해 병합된 스타일 데이터 (127) 에 기초하여 인코딩된 데이터 (193) 를 프로세싱한다.
- [0049] 도 3 을 참조하면, 제어 파라미터 추정기 (310) 및 손실 계산기 (312) 를 통해, 스피치 합성기 (130) 가 조정기 (314) 에 커플링되는 스피치 생성기 (120) 의 특정한 구현이 나타나 있다.
- [0050] 제어 파라미터 추정기 (310) 는 합성된 스피치 신호 (131) 를 프로세싱하여 하나 이상의 추정된 제어 파라미터들 (311) 을 생성한다. 예를 들어, 제어 파라미터 추정기 (310) 는 합성된 스피치 신호 (131) 에 대응하는 추정된 사람, 추정된 감정, 스피치의 추정된 레이트, 또는 이들의 조합을 결정한다.
- [0051] 특정한 예에서, 제어 파라미터 추정기 (310) 는 합성된 스피치 신호 (131) 를 도 1 의 사람들의 세트 (151) 의 각각과 연관된 스피치 모델들과 비교함으로써 모델 매치 스코어들을 생성한다. 제어 파라미터 추정기 (310) 는 합성된 스피치 신호 (131) 와 가장 근접하게 매칭되는 특정한 스피치 모델에 대응하는 모델 매치 스코어들로부터 모델 매치 스코어 (예를 들어, 가장 높은 매치 스코어) 를 선택한다. 제어 파라미터 추정기 (310) 는, 특정한 스피치 모델이 특정한 사람의 스피치를 표현한다고 결정하는 것에 응답하여, 특정한 사람을 추정된 사람으로서 표시하는 하나 이상의 추정된 제어 파라미터들 (311) 을 생성한다.
- [0052] 특정한 양태에서, 제어 파라미터 추정기 (310) 는 감정 검출 기법들을 사용하여 합성된 스피치 신호 (131) 에 의해 표시된 추정된 감정을 검출한다. 예를 들어, 제어 파라미터 추정기 (310) 는 추정된 감정을 나타내는 출력 (예를 들어, 분류기 출력 또는 임베딩) 을 생성하기 위해 합성된 스피치 신호 (131) 를 나타내는 입력 임

베딩을 프로세싱하도록 구성되는 뉴럴 네트워크를 포함한다.

- [0053] 특정한 양태에서, 제어 파라미터 추정기 (310) 는 오디오 분석 기법들을 사용하여 합성된 스피치 신호 (131) 에서 검출된 스피치의 평균 (예를 들어, 평균, 중간값, 또는 모드) 레이트를 결정하고, 스피치의 추정된 레이트로서 스피치의 평균 레이트를 표시하는 하나 이상의 추정된 제어 파라미터들 (311) 을 생성한다. 제어 파라미터 추정기 (310) 는 하나 이상의 추정된 제어 파라미터들 (311) 을 손실 계산기 (312) 에 제공한다.
- [0054] 손실 계산기 (312) 는 하나 이상의 제어 파라미터들 (125) 과 추정된 제어 파라미터들 (311) 의 비교에 기초하여 손실 메트릭 (313) 을 생성한다. 예를 들어, 손실 메트릭 (313) 은 사람 손실 메트릭, 감정 손실 메트릭, 스피치 손실 레이트 메트릭, 또는 이들의 조합을 포함한다.
- [0055] 특정한 양태에서, 타겟 사람 (143) 은 벡터 공간에서 제 1 사람 임베딩 벡터에 의해 표현되고, 추정된 사람은 벡터 공간에서 제 2 사람 임베딩 벡터에 의해 표현된다. 예를 들어, 벡터 공간의 차원들은 다양한 스피치 특성들 (예를 들어, 피치, 톤 등) 을 표현한다. 제어 파라미터 추정기 (310) 는 제 1 사람 임베딩 벡터와 제 2 사람 임베딩 벡터 사이의 벡터 거리에 기초하여 사람 손실 메트릭을 결정한다.
- [0056] 특정한 양태에서, 타겟 감정 (145) 은 벡터 공간에서 제 1 감정 임베딩 벡터에 의해 표현되고, 추정된 감정은 벡터 공간에서 제 2 감정 임베딩 벡터에 의해 표현된다. 예를 들어, 벡터 공간은 감정 차트를 나타낸다. 제어 파라미터 추정기 (310) 는 제 1 감정 임베딩 벡터와 제 2 감정 임베딩 벡터 사이의 벡터 거리에 기초하여 감정 손실 메트릭을 결정한다.
- [0057] 특정한 양태에서, 제어 파라미터 추정기 (310) 는 타겟 스피치 레이트 (147) 와 스피치의 추정된 레이트 사이의 차이에 기초하여 스피드 손실 메트릭의 레이트를 결정한다 (예를 들어, 스피치 손실 메트릭의 레이트 = 타겟 스피치 레이트 (147) 마이너스 스피치의 추정된 레이트).
- [0058] 손실 계산기 (312) 는 손실 메트릭 (313)(예를 들어, 사람 손실 메트릭, 감정 손실 메트릭, 스피치 손실 메트릭의 레이트, 또는 이들의 조합) 을 조정기 (314) 에 제공한다. 조정기 (314) 는 손실 메트릭 (313) 이 손실 임계치보다 큰 손실을 표시한다고 결정하는 것에 응답하여, 입력 표현들의 후속 프로세싱을 위해 손실 메트릭 (313) 을 감소시키도록, 멀티-인코더 (192), 사람 인코더 (206), 감정 인코더 (208), 스피치 레이트 인코더 (210), 인코더 프리-네트 (212), 디코더 (194) 의 디코더 네트워크, 또는 이들의 조합의 하나 이상의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 을 훈련하기 위해 머신 러닝 (예를 들어, 뉴럴 네트워크 훈련) 기법들을 사용한다.
- [0059] 특정한 양태에서, 조정기 (314) 는 사람 손실 메트릭이 사람 손실 임계치보다 큰 손실을 표시한다고 결정하는 것에 응답하여, 사람 인코더 (206), 인코더 프리-네트 (212), 도 4a 의 사람 인코더 (420), 마스킹된 멀티-헤드 어텐션 (466), 멀티-헤드 어텐션 (478), 또는 이들의 조합의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 을 훈련한다.
- [0060] 특정한 양태에서, 조정기 (314) 는, 감정 손실 메트릭이 감정 손실 임계치보다 큰 손실을 표시한다고 결정하는 것에 응답하여, 감정 인코더 (208), 인코더 프리-네트 (212), 도 4a 의 감정 인코더 (430), 마스킹된 멀티-헤드 어텐션 (466), 멀티-헤드 어텐션 (480), 또는 이들의 조합의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 을 훈련한다.
- [0061] 특정한 양태에서, 조정기 (314) 는 스피치 손실 메트릭이 스피치 손실 임계치의 레이트보다 큰 손실을 표시한다고 결정하는 것에 응답하여, 스피치 레이트 인코더 (210), 인코더 프리-네트 (212), 도 4a 의 레이트 인코더 (440), 마스킹된 멀티-헤드 어텐션 (466), 멀티-헤드 어텐션 (482), 또는 이들의 조합의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 을 훈련한다.
- [0062] 따라서, 조정기 (314) 는 스피치 생성기 (120) 의 훈련이 타겟 스피치 특성들 (141) 을 매칭하는데 있어서 합성 스피치 생성의 정확도를 개선하는 것을 가능하게 한다. 예를 들어, 스피치 생성기 (120) 의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 은 합성 스피치 생성의 정확도를 개선하기 위해 사용자 입력과 독립적으로 조정기 (314) 에 의해 자동으로 튜닝될 수 있다.
- [0063] 도 4a 를 참조하면, 스피치 생성기 (120) 의 특정한 구현이 나타나 있다. 멀티-인코더 (192) 는 입력 표현 인코더 (450) 및 각각의 스피치 특성에 대한 스피치 특성 인코더를 포함한다. 예를 들어, 멀티-인코더 (192) 는 사람 인코더 (420), 감정 인코더 (430), 레이트 인코더 (440), 또는 이들의 조합을 포함한다.
- [0064] 멀티-인코더 (192) 의 각각의 인코더는 멀티-인코더 (192) 의 다른 인코더들과 독립적으로, 도 5 를 참조하여

추가로 설명된 바와 같이, 입력 데이터를 프로세싱함으로써 인코딩된 데이터를 생성하도록 구성된다. 예를 들어, 입력 표현 인코더 (450) 는 제 1 인코딩된 데이터 (458) 를 생성하기 위해 입력 표현 (123) 을 프로세싱하도록 구성된다. 사람 인코더 (420) 는 사람 인코딩된 데이터 (428) 를 생성하기 위해 병합된 스타일 데이터 (127) 를 프로세싱하도록 구성된다. 감정 인코더 (430) 는 감정 인코딩된 데이터 (438) 를 생성하기 위해 병합된 스타일 데이터 (127) 를 프로세싱하도록 구성된다. 레이트 인코더 (440) 는 스피치 레이트 인코딩된 데이터 (448) 를 생성하기 위해 병합된 스타일 데이터 (127) 를 프로세싱하도록 구성된다. 멀티-인코더 (192) 의 인코더들의 각각은 멀티-인코더 (192) 의 다른 인코더들과 독립적으로 동작하기 때문에, 인코더들의 각각은 도 3 을 참조하여 설명된 바와 같이, 합성된 스피치 생성의 정확도를 개선하기 위해 대응하는 스피치 특성 손실 메트릭에 기초하여 조정기 (314) 에 의해 구체적으로 튜닝될 수 있다.

[0065] 멀티-인코더 (192) 의 각각의 인코더는 도 5 를 참조하여 추가로 설명된 바와 같이, 입력 스테이지, 하나 이상의 인코딩 계층들, 및 정규화 계층 (계층 기준; layer norm) 을 포함한다. 예를 들어, 입력 표현 인코더 (450) 는 입력 스테이지 (452), 하나 이상의 인코딩 계층들 (454), 및 계층 기준 (456) 을 포함한다. 사람 인코더 (420) 는 입력 스테이지 (422), 하나 이상의 인코딩 계층들 (424), 및 계층 기준 (426) 을 포함한다. 감정 인코더 (430) 는 입력 스테이지 (432), 하나 이상의 인코딩 계층들 (434), 및 계층 기준 (436) 을 포함한다. 레이트 인코더 (440) 는 입력 스테이지 (442), 하나 이상의 인코딩 계층들 (444), 및 계층 기준 (446) 을 포함한다.

[0066] 일부 구현들에서, 인코더들 (420, 430, 440 및 450) 의 각각은 도 5 에 도시된 인코더 (500) 를 참조하여 설명된 바와 같이 구성되고 동작한다. 도 5 에서, 인코더 (500) 는 입력 스테이지 (522), 하나 이상의 인코딩 계층들 (554), 및 계층 기준 (526) 을 포함한다. 하나 이상의 인코딩 계층들 (554) 의 각각의 인코딩 계층은 어텐션 계층 및 피드 포워드 계층을 포함한다. 어텐션 계층은 멀티-헤드 어텐션 (564) 과 같은 어텐션 네트워크를 포함한다. 피드 포워드 계층은 피드 포워드 (570)(예를 들어, 완전히 연결된 피드 포워드 뉴럴 네트워크) 와 같은 피드 포워드 뉴럴 네트워크를 포함한다. 특정한 예에서, 어텐션 계층은 멀티-헤드 어텐션 (564) 을 통해 결합기 (566) 에 커플링된 계층 기준 (562) 을 포함한다. 피드 포워드 계층은 피드 포워드 (570) 를 통해 결합기 (572) 에 커플링된 계층 기준 (568) 을 포함한다. 어텐션 계층은 피드 포워드 계층에 커플링된다. 예를 들어, 결합기 (566) 는 계층 기준 (568) 및 결합기 (572) 에 커플링된다.

[0067] 단일 인코딩 계층을 포함하는 하나 이상의 인코딩 계층들 (554) 은 예시적인 예로서 제공된다. 다른 예들에서, 하나 이상의 인코딩 계층들 (554) 은 초기 인코딩 계층의 어텐션 계층 (예를 들어, 계층 기준 (562)) 에 커플링된 입력 스테이지 (522) 의 출력, 후속 인코딩 계층의 어텐션 계층 (예를 들어, 계층 기준 (562)) 에 커플링된 각각의 이전 인코딩 계층의 피드 포워드 계층 (예를 들어, 결합기 (572)), 및 계층 기준 (526) 에 커플링된 마지막 인코딩 계층의 피드 포워드 계층 (예를 들어, 결합기 (572)) 를 갖는 다중 인코딩 계층들을 포함한다.

[0068] 입력 스테이지 (522) 는 입력 데이터 (521) 를 수신한다. 특정한 예에서, 인코더 (500) 는 도 4 의 입력 표현 인코더 (450) 에 대응한다. 예를 들어, 입력 데이터 (521) 는 입력 표현 (123) 에 대응하고, 입력 스테이지 (522) 는 입력 스테이지 (452) 에 대응하고, 하나 이상의 인코딩 계층들 (554) 은 하나 이상의 인코딩 계층들 (454) 에 대응하고, 계층 기준 (526) 은 계층 기준 (456) 에 대응한다. 특정한 구현에서, 입력 스테이지 (452) 는 출력을 생성하기 위해 입력 표현 (123) 을 프로세싱하는 인코더 프리-네트 (예를 들어, CNN, 선형 프로젝션 계층, 또는 양자 모두) 을 포함한다. 예를 들어, 입력 표현 인코더 (450) 의 인코더 프리-네트는 컨볼루션을 부가하기 위해 사용된다. 입력 스테이지 (452) 는 인코더 프리-네트의 출력을 프로세싱하도록 구성된다. 예를 들어, 입력 스테이지 (452) 는 인코더 프리-네트의 출력을 포지션 인코딩 데이터와 결합함으로써 출력을 생성하도록 구성된다.

[0069] 특정한 양태에서, 인코더 (500) 는 사람 인코더 (420), 감정 인코더 (430), 또는 레이트 인코더 (440) 에 대응하고, 입력 데이터 (521) 는 병합된 스타일 데이터 (127) 에 대응하며, 입력 스테이지 (522) 는 병합된 스타일 데이터 (127) 를 프로세싱함으로써 출력을 생성하도록 구성된다. 예를 들어, 입력 스테이지 (522) 는 병합된 스타일 데이터 (127) 를 포지션 인코딩 데이터와 결합함으로써 출력을 생성하도록 구성된다.

[0070] 특정한 예에서, 인코더 (500) 는 사람 인코더 (420) 에 대응한다. 예시를 위해, 입력 스테이지 (522) 는 입력 스테이지 (422) 에 대응하고, 하나 이상의 인코딩 계층들 (554) 은 하나 이상의 인코딩 계층들 (424) 에 대응하고, 계층 기준 (526) 은 계층 기준 (426) 에 대응한다. 특정한 예에서, 인코더 (500) 는 감정 인코더 (430) 에 대응한다. 예시를 위해, 입력 스테이지 (522) 는 입력 스테이지 (432) 에 대응하고, 하나 이상의

인코딩 계층들 (554) 은 하나 이상의 인코딩 계층들 (434) 에 대응하고, 계층 기준 (526) 은 계층 기준 (436) 에 대응한다. 특정한 예에서, 인코더 (500) 는 레이트 인코더 (440) 에 대응한다. 예시를 위해, 입력 스테이지 (522) 는 입력 스테이지 (442) 에 대응하고, 하나 이상의 인코딩 계층들 (554) 은 하나 이상의 인코딩 계층들 (444) 에 대응하고, 계층 기준 (526) 은 계층 기준 (446) 에 대응한다.

[0071] 입력 스테이지 (522) 의 출력은 계층 기준 (562) 에 의한 정규화에 후속하여 멀티-헤드 어텐션 (564) 에 제공된다. 멀티-헤드 어텐션 (564) 은 상이한 어텐션 헤드들을 사용하여 상이한 양태들로부터 컨텍스트 벡터를 구축할 수 있다. 예를 들어, 멀티-헤드 어텐션 (564) 은 멀티-헤드 어텐션 (564) 에 대한 입력을 병렬로 프로세싱하는 어텐션 헤드들을 포함한다. 예시를 위해, 입력은 제 1 행렬, 제 2 행렬 및 제 3 행렬에 의해 승산되어 제 1 쿼리 벡터, 제 1 키 벡터 및 제 1 값 벡터를 각각 생성한다. 제 1 쿼리 벡터, 제 1 키 벡터, 및 제 1 값 벡터는 제 1 어텐션 헤드에 의해 프로세싱된다. 입력은 제 4 행렬, 제 5 행렬 및 제 6 행렬에 의해 승산되어 제 2 쿼리 벡터, 제 2 키 벡터 및 제 2 값 벡터를 각각 생성한다. 제 2 쿼리 벡터, 제 2 키 벡터 및 제 2 값 벡터는 제 1 쿼리 벡터, 제 1 키 벡터 및 제 1 값 벡터를 프로세싱하는 제 1 어텐션 헤드와 병렬로 또는 동시에 제 2 어텐션 헤드에 의해 프로세싱된다.

[0072] 특정한 양태에서, 어텐션 헤드의 출력은 다음의 식에 대응한다:

$$Z = \text{softmax}\left(\frac{QxK^T}{\sqrt{d_k}}\right)V,$$

[0073] 식중, Z 는 어텐션 헤드의 출력에 대응하고, Q 는 쿼리 벡터에 대응하고, x 는 승산 연산자에 대응하고, K 는 키 벡터에 대응하고, V 는 값 벡터에 대응하고, d_k 는 키 벡터들의 차원에 대응하며, 소프트맥스는 정규화 연산에 대응한다.

[0075] 어텐션 헤드들의 독립적인 출력들은 멀티-헤드 어텐션 (564) 의 출력을 생성하도록 연결되고 선형으로 변환된다. 결합기 (566) 는 계층 기준 (562) 에 대한 입력 및 멀티-헤드 어텐션 (564) 의 출력을 결합함으로써 출력을 생성한다.

[0076] 결합기 (566) 의 출력은 계층 기준 (568) 에 의한 정규화에 후속하여, 피드 포워드 (570)(예를 들어, 완전히 연결된 피드 포워드 뉴럴 네트워크) 에 제공된다. 특정한 예에서, 피드 포워드 (570) 는 정류된 선형 유닛 (ReLU) 계층을 통해 제 2 선형 변환 계층에 커플링된 제 1 선형 변환 계층을 포함한다. 피드 포워드 (570) 는 결합기 (566) 의 출력을 프로세싱함으로써 출력을 생성한다.

[0077] 결합기 (572) 는 결합기 (566) 의 출력을 피드 포워드 (570) 의 출력과 결합함으로써 출력을 생성한다. 특정한 양태에서, 계층 기준 (526) 은 결합기 (572) 의 (예를 들어, 하나 이상의 인코딩 계층들 (554) 중 마지막 인코딩 계층의) 출력에 정규화를 적용함으로써 인코딩된 데이터 (528) 를 생성한다. 예를 들어, 정규화를 적용하는 것은 결합기 (572) 의 출력 (예를 들어, 출력 벡터) 의 각각의 값을 특정한 범위 내에 있도록 조정하는 것을 포함한다.

[0078] 도 4a 로 돌아가면, 입력 표현 인코더 (450) 는 제 1 인코딩된 데이터 (458) 를 생성하기 위해 하나 이상의 제어 파라미터들 (125) 과 독립적으로 입력 표현 (123) 을 인코딩하도록 구성된다. 사람 인코더 (420), 감정 인코더 (430), 및 레이트 인코더 (440) 의 각각은 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들 (125) 에 기초하여 입력 표현 (123) 을 인코딩하도록 구성된다. 예를 들어, 하나 이상의 제어 파라미터 (125) 에 기초하여 입력 표현 (123) 을 인코딩하는 것은 병합된 스타일 데이터 (127) 를 인코딩하는 것을 포함한다. 예시를 위해, 사람 인코더 (420) 는 병합된 스타일 데이터 (127) 를 인코딩함으로써 사람 인코딩된 데이터 (428) 를 생성한다. 감정 인코더 (430) 는 병합된 스타일 데이터 (127) 를 인코딩함으로써 감정 인코딩된 데이터 (438) 를 생성한다. 레이트 인코더 (440) 는 병합된 스타일 데이터 (127) 를 인코딩함으로써 스피치 레이트 인코딩된 데이터 (448) 를 생성한다.

[0079] 특정한 양태에서, 사람 인코더 (420), 감정 인코더 (430), 및 레이트 인코더 (440) 는 대응하는 멀티-헤드 어텐션 (564) 에 의해 사용된 가중치들 또는 행렬들에서 서로 상이하다. 예를 들어, 사람 인코더 (420) 의 멀티-헤드 어텐션 (564) 은 쿼리 벡터들, 키 벡터들 및 값 벡터들을 생성하기 위해 가중치들의 제 1 세트를 표시하는 행렬들의 제 1 세트를 사용한다. 감정 인코더 (430) 의 멀티-헤드 어텐션 (564) 은 쿼리 벡터들, 키 벡터들, 및 값 벡터들을 생성하기 위해 가중치들의 제 2 세트를 표시하는 행렬들의 제 2 세트를 사용한다. 유사하게, 레이트 인코더 (440) 의 멀티-헤드 어텐션 (564) 은 쿼리 벡터들, 키 벡터들, 및 값 벡터들을 생성하기

위해 가중치들의 제 3 세트를 표시하는 행렬들의 제 3 세트를 사용한다.

- [0080] 멀티-인코더 (192) 는 인코딩된 데이터 (193) 를 디코더 (194) 에 제공한다. 예를 들어, 인코딩된 데이터 (193) 는 하나 이상의 제어 파라미터들 (125) 과는 독립적인 제 1 인코딩된 데이터 (458) 를 포함한다. 특정한 양태에서, 인코딩된 데이터 (193) 는 또한 하나 이상의 제어 파라미터들 (125) 에 기초하는 제 2 인코딩된 데이터 (예를 들어, 사람 인코딩된 데이터 (428), 감정 인코딩된 데이터 (438), 스피치 레이트 인코딩된 데이터 (448), 또는 이들의 조합) 를 포함한다.
- [0081] 디코더 (194) 는 디코더 프리-네트 (460), 결합기 (462), 하나 이상의 디코딩 계층들 (498), 계층 기준 (492), 선형 변환 계층 (선형)(494), 또는 이들의 조합을 포함한다. 하나 이상의 디코딩 계층들 (498) 의 각각의 디코딩 계층은 마스킹된 어텐션 계층, 어텐션 계층, 및 피드 포워드 계층을 포함한다. 예를 들어, 마스킹된 어텐션 계층은 마스킹된 멀티-헤드 어텐션 (466)(예를 들어, 마스킹된 디코더 어텐션 네트워크) 을 통해 계층 기준 (468), 계층 기준 (470), 계층 기준 (472), 계층 기준 (474), 또는 이들의 조합과 같은 복수의 정규화 계층들 (계층 기준들) 에 커플링된 계층 기준 (464) 을 포함한다. 어텐션 계층은 결합기 (486) 를 통해 계층 기준 (488) 에 커플링된 복수의 디코더 어텐션 네트워크들을 포함한다. 예를 들어, 복수의 디코더 어텐션 네트워크들 (예를 들어, 디코더 멀티-헤드 어텐션 네트워크들) 은 멀티-헤드 어텐션 (478), 멀티-헤드 어텐션 (480), 멀티-헤드 어텐션 (482), 멀티-헤드 어텐션 (484), 또는 이들의 조합을 포함한다. 피드 포워드 계층은 피드 포워드 (490)(예를 들어, 완전히 연결된 피드 포워드 뉴럴 네트워크) 를 포함한다.
- [0082] 마스킹된 어텐션 계층은 어텐션 계층에 커플링된다. 예를 들어, 마스킹된 어텐션 계층의 복수의 계층 기준들의 각각은 어텐션 계층의 대응하는 멀티-헤드 어텐션 네트워크에 커플링된다. 예를 들어, 계층 기준 (468), 계층 기준 (470), 계층 기준 (472), 및 계층 기준 (474) 은 각각, 멀티-헤드 어텐션 (478), 멀티-헤드 어텐션 (480), 멀티-헤드 어텐션 (482), 및 멀티-헤드 어텐션 (484) 에 커플링된다. 마스킹된 멀티-헤드 어텐션 (466) 은 결합기 (486) 를 통해 계층 기준 (488) 에 커플링된다. 어텐션 계층은 피드 포워드 계층에 커플링된다. 예를 들어, 계층 기준 (488) 은 피드 포워드 (490) 에 커플링된다.
- [0083] 결합기 (462) 는 하나 이상의 디코딩 계층들 (498) 중 디코딩 계층에 커플링된다. 예를 들어, 결합기 (462) 는 디코딩 계층의 계층 기준 (464) 에 커플링된다. 하나 이상의 디코딩 계층들 (498) 의 디코딩 계층은 계층 기준 (492) 에 커플링된다. 예를 들어, 디코딩 계층의 피드 포워드 (490) 는 계층 기준 (492) 에 커플링된다.
- [0084] 특정한 양태에서, 디코더 프리-네트 (460) 은 이전 시간 단계 동안 디코더 (194) 에 의해 생성된 출력 스펙트럼 데이터 (129) 에 대응하는 피드백 출력 스펙트럼 데이터 (429) 를 수신한다. 디코더 프리-네트 (460) 는 출력을 생성하기 위해 피드백 출력 스펙트럼 데이터 (429) 를 프로세싱하는 뉴럴 네트워크 (예를 들어, CNN) 를 포함한다. 특정한 양태에서, 디코더 프리-네트 (460) 는 컨볼루션을 부가한다. 특정한 양태에서, 결합기 (462) 는 디코더 프리-네트 (460) 의 출력과 포지션 인코딩 데이터 (461) 를 결합함으로써 출력을 생성한다.
- [0085] 결합기 (462) 의 출력은 하나 이상의 디코딩 계층들 (498) 중 디코딩 계층에 의해 프로세싱된다. 예를 들어, 결합기 (462) 의 출력은 디코딩 계층의 계층 기준 (464) 에 의한 정규화에 후속하여, 마스킹된 멀티-헤드 어텐션 (466) 에 제공된다. 특정한 양태에서, 마스킹된 멀티-헤드 어텐션 (466) 은 마스킹된 멀티-헤드 어텐션 (466) 에 대한 입력에서 향후 포지션들을 마스킹한다. 마스킹된 멀티-헤드 어텐션 (466) 은 마스킹된 멀티-헤드 어텐션 (466) 에 대한 입력의 마스킹된 버전으로부터 쿼리 벡터들, 키 벡터들, 및 값 벡터들을 생성한다. 마스킹된 멀티-헤드 어텐션 (466) 의 각각의 어텐션 헤드는 출력을 생성하기 위해 쿼리 벡터, 키 벡터, 및 값 벡터를 프로세싱한다. 마스킹된 멀티-헤드 어텐션 (466) 의 어텐션 헤드들의 독립적인 출력들은 마스킹된 멀티-헤드 어텐션 (466) 의 출력을 생성하도록 연결되고 선형으로 변환된다.
- [0086] 마스킹된 멀티-헤드 어텐션 (466) 의 출력은 결합기 (486) 에 제공되고, 정규화에 후속하여, 디코딩 계층의 복수의 멀티-헤드 어텐션에 제공된다. 예를 들어, 마스킹된 멀티-헤드 어텐션 (466) 의 출력은 계층 기준 (468), 계층 기준 (470), 계층 기준 (472), 및 계층 기준 (474) 에 의한 정규화에 후속하여, 각각 멀티-헤드 어텐션 (478), 멀티-헤드 어텐션 (480), 멀티-헤드 어텐션 (482), 및 멀티-헤드 어텐션 (484) 에 제공된다.
- [0087] 디코딩 계층의 복수의 멀티-헤드 어텐션들의 각각은 마스킹된 멀티-헤드 어텐션 (466) 의 출력의 정규화된 버전 및 멀티-인코더 (192) 의 대응하는 인코더로부터 수신된 인코딩된 데이터를 프로세싱하도록 구성된다. 예를 들어, 디코딩 계층의 복수의 멀티-헤드 어텐션들의 각각은 마스킹된 멀티-헤드 어텐션 (466) 의 출력의 정규화된 버전에 기초하는 쿼리 벡터들을 프로세싱 (예를 들어, 생성 및 프로세싱) 하고, 멀티-인코더 (192) 의 대응

하는 인코더로부터 수신된 인코딩된 데이터에 기초하는 키 벡터들 및 값 벡터들을 프로세싱 (예를 들어, 생성 및 프로세싱) 한다. 예시를 위해, 멀티-헤드 어텐션 (484) 은 계층 기준 (474) 의 출력에 기초하는 쿼리 벡터들을 프로세싱하고, 제 1 인코딩된 데이터 (458) 에 기초하는 키 벡터들 및 값 벡터들을 프로세싱하도록 구성된다. 멀티-헤드 어텐션 (478), 멀티-헤드 어텐션 (480), 및 멀티-헤드 어텐션 (482) 은, 각각 사람 인코딩된 데이터 (428), 감정 인코딩된 데이터 (438), 및 스피치 레이트 인코딩된 데이터 (448), 및 각각 계층 기준 (468), 계층 기준 (470), 및 계층 기준 (472) 의 출력을 프로세싱하도록 구성된다. 예를 들어, 멀티-헤드 어텐션 (478) 은 계층 기준 (468) 의 출력에 기초하는 쿼리 벡터들을 프로세싱하고, 사람 인코딩된 데이터 (428) 에 기초하는 키 벡터들 및 값 벡터들을 프로세싱한다.

[0088] 결합기 (486) 는 마스킹된 멀티-헤드 어텐션 (466) 의 출력을 멀티-헤드 어텐션 (478) 의 출력, 멀티-헤드 어텐션 (480) 의 출력, 멀티-헤드 어텐션 (482) 의 출력, 멀티-헤드 어텐션 (484) 의 출력, 또는 이들의 조합과 결합함으로써 출력을 생성한다. 결합기 (486) 의 출력은 계층 기준 (488) 에 의한 정규화에 후속하여, 디코딩 계층의 피드 포워드 (490) 에 제공된다.

[0089] 하나 이상의 디코딩 계층들 (498) 중 특정한 디코딩 계층의 피드 포워드 (490) 의 출력은, 계층 기준 (492) 에 의한 정규화 및 계층 기준 (492) 에 의한 선형 변환에 후속하여, 출력 스펙트럼 데이터 (129) 로서 스피치 합성기 (130) 에 제공된다. 특정한 양태에서, 스피치 합성기 (130) 는 합성된 스피치 신호 (131) 를 생성하기 위해, 출력 스펙트럼 데이터 (129), 인코딩된 데이터 (193), 또는 이들의 조합에 기초하여 스피치 합성을 수행한다. 출력 스펙트럼 데이터 (129) 는 후속 시간 단계 동안 피드백 출력 스펙트럼 데이터 (429) 로서 디코더 프리-네트 (460) 에 제공된다.

[0090] 단일 디코딩 계층을 포함하는 하나 이상의 디코딩 계층들 (498) 은 예시적인 예로서 제공된다. 다른 예들에서, 하나 이상의 디코딩 계층들 (498) 은 초기 디코딩 계층의 계층 기준 (464) 에 커플링된 결합기 (462) 의 출력, 후속 인코딩 계층의 계층 기준 (464) 에 커플링된 각각의 이전 디코딩 계층의 피드 포워드 (490), 및 계층 기준 (492) 에 커플링된 마지막 디코딩 계층의 피드 포워드 (490) 를 갖는 다중 디코딩 계층들을 포함한다. 멀티-인코더 (192) 의 마지막 인코딩 계층은 인코딩된 데이터 (193) 를 하나 이상의 디코딩 계층들 (498) 각각의 복수의 멀티-헤드 어텐션들에 제공한다.

[0091] 특정한 양태에서, 도 3 의 조정기 (314) 는 손실 메트릭 (313) 에 기초하여 멀티-인코더 (192) 및 디코더 (194) 의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들) 을 조정한다. 예를 들어, 조정기 (314) 는 손실 메트릭 (313) 의 사람 손실 메트릭이 사람 손실 임계치를 만족시키지 못한다고 결정하는 것에 응답하여, 사람 인코더 (420) 의 하나 이상의 인코딩 계층들 (424), 멀티-헤드 어텐션 (478), 마스킹된 멀티-헤드 어텐션 (466), 하나 이상의 디코딩 계층들 (498) 의 피드 포워드 (490), 디코더 프리-네트 (460), 또는 이들의 조합의 뉴럴 네트워크 가중치들을 훈련한다.

[0092] 도 4b 를 참조하면, 스피치 생성기 (120) 의 특정한 구현이 나타나 있다. 멀티-인코더 (192) 는 각각의 스피치 특성에 대한 별도의 스피치 특성 대신 스피치 특성들에 대한 특성 인코더 (451) 를 포함한다. 예를 들어, 특성 인코더 (451) 는 입력 스테이지 (453), 하나 이상의 인코딩 계층들 (455), 및 계층 기준 (457) 을 포함한다. 특정한 양태에서, 도 5 의 인코더 (500) 는 특성 인코더 (451) 의 구현의 예이다. 예를 들어, 입력 스테이지 (522) 는 입력 스테이지 (453) 에 대응하고, 하나 이상의 인코딩 계층들 (554) 은 하나 이상의 인코딩 계층들 (455) 에 대응하고, 계층 기준 (526) 은 계층 기준 (457) 에 대응한다.

[0093] 특성 인코더 (451) 는 스타일 데이터 생성기 (126) 로부터 병합된 스타일 데이터 (127)(예를 들어, 입력 임베딩 벡터) 를 수신한다. 특정한 양태에서, 입력 스테이지 (453) 는 병합된 스타일 데이터 (127) 를 포지션 인코딩 데이터와 결합함으로써 출력을 생성한다. 입력 스테이지 (453) 의 출력은 도 5 의 하나 이상의 인코딩 계층들 (554) 을 참조하여 설명된 바와 같이, 하나 이상의 인코딩 계층들 (455) 에 의해 프로세싱된다. 하나 이상의 인코딩 계층들 (455) 의 출력은, 계층 기준 (457) 에 의한 정규화에 후속하여, 제 2 인코딩된 데이터 (459) 로서 디코더 (194) 에 제공된다. 예를 들어, 인코딩된 데이터 (193) 는 제 1 인코딩된 데이터 (458) 및 제 2 인코딩된 데이터 (459) 를 포함한다.

[0094] 하나 이상의 디코딩 계층들 (498) 은 정규화 계층들 (계층 규범) 을 통해 2개의 멀티-헤드 어텐션들에 커플링된 마스킹된 멀티-헤드 어텐션 (466) 을 포함한다. 예를 들어, 마스킹된 멀티-헤드 어텐션 (466) 은 계층 기준 (475) 을 통해 멀티-헤드 어텐션 (485) 에 그리고 계층 기준 (474) 을 통해 멀티-헤드 어텐션 (484) 에 커플링된다. 결합기 (486) 는 마스킹된 멀티-헤드 어텐션 (466), 멀티-헤드 어텐션 (485) 및 멀티-헤드 어텐션 (484) 에 커플링된다.

- [0095] 마스크된 멀티-헤드 어텐션 (466)의 출력은 계층 기준 (475)에 의한 정규화에 후속하여, 멀티-헤드 어텐션 (485)에 제공된다. 특성 인코더 (451)는 제 2 인코딩된 데이터 (459)를 멀티-헤드 어텐션 (485)에 제공한다. 멀티-헤드 어텐션 (485)은 마스크된 멀티-헤드 어텐션 (466)의 출력에 기초하여 쿼리 벡터들을 생성하고, 제 2 인코딩된 데이터 (459)에 기초하여 키 벡터들 및 값 벡터들을 생성한다. 멀티-헤드 어텐션 (485)의 각각의 어텐션 헤드는 출력을 생성하기 위해 쿼리 벡터, 키 벡터, 및 값 벡터를 프로세싱한다. 멀티-헤드 어텐션 (485)의 어텐션 헤드들의 각각의 출력들은 멀티-헤드 어텐션 (485)의 출력을 생성하도록 연결되고 선형으로 변환된다. 결합기 (486)는 마스크된 멀티-헤드 어텐션 (466)의 출력을 멀티-헤드 어텐션 (485)의 출력, 멀티-헤드 어텐션 (484)의 출력, 또는 양자 모두와 결합함으로써 출력을 생성한다.
- [0096] 특정한 양태에서, 도 3의 조정기 (314)는 손실 메트릭 (313)에 기초하여 멀티-인코더 (192) 및 디코더 (194)의 구성 파라미터들 (예를 들어, 뉴럴 네트워크 가중치들)을 조정한다. 예를 들어, 조정기 (314)는 손실 메트릭 (313)의 사람 손실 메트릭, 감정 손실 메트릭, 또는 스피치 손실 레이트 메트릭 중 임의의 것이 대응하는 손실 임계치를 만족시키지 못한다고 결정하는 것에 응답하여, 특성 인코더 (451)의 하나 이상의 인코딩 계층들 (455), 멀티-헤드 어텐션 (485), 마스크된 멀티-헤드 어텐션 (466), 하나 이상의 디코딩 계층들 (498)의 피드 포워드 (490), 디코더 프리-네트 (460), 또는 이들의 조합의 뉴럴 네트워크 가중치들을 훈련한다.
- [0097] 도 6을 참조하면, 도 1의 시스템 (100)의 컴포넌트들의 동작들의 예시적인 양태가 나타나 있으며 일반적으로 600으로 지정된다. 입력 신호 (121)는 복수의 오디오 샘플들 (602), 예를 들어, 오디오 샘플들 (602A), 오디오 샘플들 (602B), N 번째 오디오 샘플들 (602N)(여기서, N은 2보다 큰 정수)과 같은 오디오 샘플들의 하나 이상의 부가 세트들, 또는 이들의 조합을 포함한다. 오디오 샘플들 (602)은 특정한 사람 (예를 들어, 사람 (180))의 스피치, 특정한 감정 (예를 들어, 중립 또는 지루함), 스피치의 특정한 레이트 (예를 들어, 중간 속도), 또는 이들의 조합을 나타낸다.
- [0098] 스피치 생성기 (120)는 하나 이상의 제어 파라미터들 (621)을 수신 (예를 들어, 액세스)한다. 하나 이상의 제어 파라미터들 (621)은 타겟 사람 (623)(예를 들어, 사람 (180)), 타겟 감정 (625)(예를 들어, 흥분), 타겟 스피치 레이트 (627) (예를 들어, 스피치의 빠른 레이트), 또는 이들의 조합을 표시한다. 하나 이상의 제어 파라미터들 (621)은 입력 신호 (121)에 기초하여 스피치 생성기 (120)에 의해 생성될 합성된 스피치 신호 (661)의 스피치 특성들을 표시한다. 특정한 예에서, 타겟 사람 (623), 타겟 감정 (625) 및 타겟 스피치 레이트 (627) 중 적어도 하나는 각각 특정한 사람, 특정한 감정 및 스피치의 특정한 레이트와 상이하다. 스피치 생성기 (120)는 합성된 스피치 신호 (661)를 생성하기 위해, 도 1을 참조하여 설명된 바와 같이, 하나 이상의 제어 파라미터들 (621)에 기초하여 입력 신호 (121)를 프로세싱한다. 합성된 스피치 신호 (661)는 타겟 사람 (623), 타겟 감정 (625), 타겟 스피치 레이트 (627), 또는 이들의 조합에 기초하는 합성된 오디오 샘플들 (629)을 포함한다. 예를 들어, 합성된 오디오 샘플들 (629)은 타겟 사람 (623)(예를 들어, 사람 (180))과 유사한 스피치 특성들, 타겟 감정 (625)(예를 들어, 흥분)과 연관된 유사한 스피치 특성들 (예를 들어, 피치 또는 톤), 타겟 스피치 레이트 (627)(예를 들어, 스피치의 빠른 레이트), 또는 이들의 조합을 갖는 스피치를 나타낸다. 특정한 양태에서, 스피치 생성기 (120)(또는 디바이스 (102)의 다른 컴포넌트)는 오디오 샘플들 (602) 및 합성된 오디오 샘플들 (629)을 훈련 데이터세트 (665)에 추가한다.
- [0099] 특정한 양태에서, 스피치 생성기 (120)는 동일한 입력 신호 (121)로부터, 다양한 제어 파라미터들에 대응하는 다양한 스피치 특성들을 갖는 오디오 샘플들의 다중 세트들을 생성한다. 예를 들어, 스피치 생성기 (120)는 타겟 사람 (633)(예를 들어, 사람 (180)), 타겟 감정 (635) (예를 들어, 피곤), 타겟 스피치 레이트 (637) (예를 들어, 느린 스피치 레이트), 또는 이들의 조합을 표시하는 하나 이상의 제어 파라미터들 (631)을 수신 (예를 들어, 액세스)한다. 스피치 생성기 (120)는 합성된 스피치 신호 (663)를 생성하기 위해, 도 1을 참조하여 설명된 바와 같이, 하나 이상의 제어 파라미터들 (631)에 기초하여 입력 신호 (121)를 프로세싱한다. 합성된 스피치 신호 (663)는 타겟 사람 (633) (예를 들어, 사람 (180)), 타겟 감정 (635) (예를 들어, 피곤), 타겟 스피치 레이트 (637)(예를 들어, 스피치의 느린 레이트), 또는 이들의 조합에 기초하는 합성된 오디오 샘플들 (639)을 포함한다. 특정한 양태에서, 스피치 생성기 (120)(또는 디바이스 (102)의 다른 컴포넌트)는 합성된 오디오 샘플들 (639)을 훈련 데이터세트 (665)에 추가한다.
- [0100] 네트워크 훈련기 (670)는 훈련 데이터세트 (665)에 기초하여 스피치 인식 네트워크 (672)를 훈련하기 위해 머신 러닝 기법들을 사용한다. (오디오 샘플들 (602)에 추가하여) 훈련 데이터세트 (665)에 합성된 오디오 샘플들 (629) 및 합성된 오디오 샘플들 (639)을 포함하는 것은 스피치를 인식함에 있어서 스피치 인식 네트워크 (672)의 정확도를 개선한다. 예를 들어, 스피치 인식 네트워크 (672)(합성된 오디오 샘플들 (629) 및 합성된 오디오 샘플들 (639)에 기초하여 훈련됨)는 입력 신호 (121)에서 사람 (180)에 의해 제공된 스피치

의 특성들 (예를 들어, 중립) 과는 상이한 특성들 (예를 들어, 빠른 또는 느린 및 피곤 또는 흥분) 을 갖는 사람 (180) 의 스피치를 인식하는데 더 강건하다.

[0101] 도 7 은 도 1 의 시스템의 컴포넌트들의 동작의 예시적인 양태의 다이어그램이다. 오디오 분석기 (122) 는 제 1 프레임 (F1)(712), 제 2 프레임 (F2)(714), 및 제 N 프레임 (FN)(716)(여기서 N 은 2 보다 큰 정수) 을 포함하는 하나 이상의 부가 프레임들로서 예시된, 입력 신호 (121) 의 연속적으로 캡처된 프레임들의 시퀀스와 같은, 오디오 데이터 샘플들의 시퀀스 (710) 를 수신하도록 구성된다. 오디오 분석기 (122) 는 제 1 세트 (722), 제 2 세트 (724), 및 제 N 세트 (726) 를 포함하는 하나 이상의 부가 세트들을 포함하는 입력 표현 (123) 의 세트들의 시퀀스와 같은, 입력 표현 데이터의 세트들의 시퀀스 (720) 를 출력하도록 구성된다.

[0102] 스타일 데이터 생성기 (126) 는 입력 표현 데이터의 세트들의 시퀀스 (720) 를 수신하고 시퀀스 (720) 및 하나 이상의 제어 파라미터들 (125) 에 기초하여 병합된 스타일 데이터의 세트들의 시퀀스 (730) 를 생성하도록 구성된다. 시퀀스 (730) 는 제 1 세트 (M1)(732), 제 2 세트 (M2)(734), 및 제 N 세트 (MN)(736) 를 포함하는 하나 이상의 부가 세트들을 포함한다. 멀티-인코더 (192) 는 시퀀스 (720) 및 시퀀스 (730) 를 수신하고 제 1 인코딩된 세트 (E1)(742), 제 2 인코딩된 세트 (E2)(744), 및 제 N 세트 (EN)(746) 를 포함하는 하나 이상의 부가 인코딩된 세트들과 같은, 데이터의 인코딩된 세트들의 시퀀스 (740) 를 생성하도록 구성된다. 디코더 (194) 는 인코딩된 세트들의 시퀀스 (740) 를 수신하고, 제 1 스펙트럼 데이터 (S1)(752), 제 2 스펙트럼 데이터 (S2)(754), 및 제 N 스펙트럼 데이터 (SN)(756) 를 포함하는 스펙트럼 데이터의 하나 이상의 부가 세트들을 포함하는 스펙트럼 데이터의 시퀀스 (750) 를 생성하도록 구성된다.

[0103] 동작 동안, 오디오 분석기 (122) 는 입력 표현 (123) 의 제 1 세트 (722) 를 생성하기 위해 제 1 프레임 (F1)(712) 을 프로세싱하고 스타일 데이터 생성기 (126) 는 병합된 스타일 데이터 (127) 의 제 1 세트 (M1)(732) 를 생성하기 위해 제 1 세트 (722) 를 프로세싱한다. 멀티-인코더 (192) 는 인코딩된 데이터 (193) 의 제 1 인코딩된 세트 (E1) 를 생성하기 위해 입력 표현 (123) 의 제 1 세트 (722) 및 병합된 스타일 데이터 (127) 의 제 1 세트 (M1)(732) 를 프로세싱한다. 디코더 (194) 는 출력 스펙트럼 데이터 (129) 의 제 1 스펙트럼 데이터 (S1)(752) 를 생성하기 위해 인코딩된 데이터 (193) 의 제 1 인코딩된 세트 (E1) 를 프로세싱한다. 후속 인코딩된 세트 (예를 들어, 제 2 인코딩된 세트 (E2)) 는 제 1 스펙트럼 데이터 (S1)(752) 에 적어도 부분적으로 기초하여 디코더 (194) 에 의해 프로세싱된다. 이러한 프로세싱은 입력 표현 (123) 의 제 N 세트 (726) 를 생성하기 위해 제 N 프레임 (716) 을 프로세싱하는 오디오 분석기 (122), 병합된 스타일 데이터 (127) 의 제 N 세트 (MN)(736) 를 생성하기 위해 입력 표현 (123) 의 제 N 세트 (726) 를 프로세싱하는 스타일 데이터 생성기 (126), 인코딩된 데이터 (193) 의 제 N 인코딩된 세트 (EN)(746) 를 생성하기 위해 입력 표현 (123) 의 제 N 세트 (726) 및 병합된 스타일 데이터 (127) 의 제 N 세트 (MN)(736) 를 프로세싱하는 멀티-인코더 (192), 및 제 N 스펙트럼 데이터 (SN)(756) 를 생성하기 위해 인코딩된 데이터 (193) 의 제 N 인코딩된 세트를 프로세싱하는 디코더 (194) 를 포함하여 계속된다. 제 N 스펙트럼 데이터 (SN)(756) 는 특징 데이터의 제 N 세트 (726) 에 기초하고 시퀀스 (720) 의 특징 데이터의 이전 세트들 중 하나 이상에 적어도 부분적으로 기초한다. 스펙트럼 데이터의 하나 이상의 이전 세트들에 기초하여 디코딩함으로써, 디코더 (194) 에 의한 합성된 스피치 생성의 정확도는 오디오 데이터의 다중 프레임들에 걸쳐 있을 수도 있는 스피치 신호들에 대해 개선될 수도 있다.

[0104] 도 8 은 하나 이상의 프로세서들 (190) 을 포함하는 집적 회로 (802) 로서의 디바이스 (102) 의 구현 (800) 을 도시한다. 집적 회로 (802) 는 또한 입력 신호 (121) 가 프로세싱을 위해 수신되는 것을 가능하게 하기 위해, 하나 이상의 버스 인터페이스들과 같은 오디오 입력 (804) 을 포함한다. 집적 회로 (802) 는 또한 합성된 스피치 신호 (131) 와 같은 출력 신호의 전송을 가능하게 하기 위해, 버스 인터페이스와 같은 오디오 출력 (806) 을 포함한다. 집적 회로 (802) 는 도 9 에 도시된 바와 같은 모바일 폰 또는 태블릿, 도 10 에 도시된 바와 같은 헤드셋, 도 11 에 도시된 바와 같은 웨어러블 전자 디바이스, 도 12 에 도시된 바와 같은 음성-제어형 스피커 시스템, 도 13 에 도시된 바와 같은 카메라, 도 14 에 도시된 바와 같은 가상 현실 헤드셋 또는 증강 현실 헤드셋, 또는 도 15 또는 도 16 에 도시된 바와 같은 차량과 같은, 마이크로폰들을 포함하는 시스템에서의 컴포넌트로서 합성된 스피치 생성의 구현을 가능하게 한다.

[0105] 도 9 는 디바이스 (102) 가 예시적이고 비제한적인 예로서 전화 또는 태블릿과 같은 모바일 디바이스 (902) 를 포함하는 구현 (900) 을 도시한다. 모바일 디바이스 (902) 는 마이크로폰 (182), 및 디스플레이 스크린 (904) 을 포함한다. 스피치 생성기 (120) 를 포함하는 프로세서 (190) 의 컴포넌트들은 모바일 디바이스 (902) 에 통합되고, 모바일 디바이스 (902) 의 사용자에게 일반적으로 보이지 않는 내부 컴포넌트들을 표시하기 위해 파선들을 사용하여 도시된다. 특정한 예에서, 스피치 생성기 (120) 는 합성된 스피치를 생성하도록 동

작하며 (또는 도 6 의 스피치 인식 네트워크 (672) 는 스피치를 인식하도록 동작함), 이는 그 후, (예를 들어, 통합된 "스마트 어시스턴트" 애플리케이션을 통해) 그래픽 사용자 인터페이스를 론칭하거나 그렇지 않으면 디스플레이 스크린 (904) 에서 사용자의 스피치와 연관된 다른 정보를 디스플레이하는 것과 같은, 모바일 디바이스 (902) 에서의 하나 이상의 동작들을 수행하도록 프로세싱된다.

[0106] 도 10 은 디바이스 (102) 가 헤드셋 디바이스 (1002) 를 포함하는 구현 (1000) 을 도시한다. 헤드셋 디바이스 (1002) 는 마이크로폰 (182) 을 포함한다. 스피치 생성기 (120) 를 포함하는 프로세서 (190) 의 컴포넌트들은 헤드셋 디바이스 (1002) 에 통합된다. 특정한 예에서, 스피치 생성기 (120) 는 합성된 스피치를 생성하도록 동작하며 (또는 도 6 의 스피치 인식 네트워크 (672) 가 스피치를 인식하도록 동작함), 이는 헤드셋 디바이스 (1002) 로 하여금 헤드셋 디바이스 (1002) 에서 하나 이상의 동작들을 수행하게 하거나, 추가 프로세싱을 위해 합성된 스피치에 대응하는 오디오 데이터를 제 2 디바이스 (미도시) 로 송신하게 하거나, 이들의 조합을 수행하게 할 수도 있다.

[0107] 도 11 은 디바이스 (102) 가 "스마트 워치" 로서 예시된 웨어러블 전자 디바이스 (1102) 를 포함하는 구현 (1100) 을 도시한다. 스피치 생성기 (120) 및 마이크로폰 (182) 은 웨어러블 전자 디바이스 (1102) 에 통합된다. 특정한 예에서, 스피치 생성기 (120) 는 합성된 스피치를 생성하도록 동작하며 (또는 도 6 의 스피치 인식 네트워크 (672) 가 스피치를 인식하도록 동작함), 이는 그 후, 그래픽 사용자 인터페이스를 론칭하거나 그렇지 않으면 웨어러블 전자 디바이스 (1102) 의 디스플레이 스크린 (1104) 에서 사용자의 스피치와 연관된 다른 정보를 디스플레이하는 것과 같은, 웨어러블 전자 디바이스 (1102) 에서의 하나 이상의 동작들을 수행하도록 프로세싱된다. 예시를 위해, 웨어러블 전자 디바이스 (1102) 는 웨어러블 전자 디바이스 (1102) 에 의해 검출된 사용자 스피치에 기초하여 통지를 디스플레이하도록 구성되는 디스플레이 스크린을 포함할 수도 있다. 특정한 예에서, 웨어러블 전자 디바이스 (1102) 는 사용자 음성 활동의 검출 또는 합성된 스피치의 생성에 응답하여 햅틱 통지 (예를 들어, 진동들) 를 제공하는 햅틱 디바이스를 포함한다. 예를 들어, 햅틱 통지는 사용자에게 의해 스피킹된 키워드의 검출을 표시하는 디스플레이된 통지를 보기 위해 사용자가 웨어러블 전자 디바이스 (1102) 를 보게 할 수 있다. 따라서, 웨어러블 전자 디바이스 (1102) 는 청각 장애가 있는 사용자 또는 헤드셋을 착용하는 사용자에게 사용자의 음성 활동이 검출됨을 알릴 수 있다.

[0108] 도 12 는 디바이스 (102) 가 무선 스피커 및 음성 활성화 디바이스 (1202) 를 포함하는 구현 (1200) 이다. 무선 스피커 및 음성 활성화 디바이스 (1202) 는 무선 네트워크 연결성을 가질 수 있고 어시스턴트 동작을 실행하도록 구성된다. 스피치 생성기 (120), 마이크로폰 (182), 또는 이들의 조합을 포함하는 프로세서 (190) 는 무선 스피커 및 음성 활성화 디바이스 (1202) 에 포함된다. 무선 스피커 및 음성 활성화 디바이스 (1202) 는 또한 스피커 (1204) 를 포함한다. 동작 동안, 구두 커맨드를 수신하고 스피치 생성기 (120) 의 동작을 통해 합성된 스피치를 생성하는 것 (또는 도 6 의 스피치 인식 네트워크 (672) 의 동작을 통해 스피치를 인식하는 것) 에 응답하여, 무선 스피커 및 음성 활성화 디바이스 (1202) 는, 예컨대 통합된 어시스턴트 애플리케이션의 실행을 통해 어시스턴트 동작들을 실행할 수 있다. 어시스턴트 동작들은 온도를 조정하는 것, 음악을 재생하는 것, 조명들을 턴온하는 것 등을 포함할 수 있다. 예를 들어, 어시스턴트 동작들은 키워드 또는 핵심 문구 (예를 들어, "안녕 어시스턴트") 후에 커맨드를 수신하는 것에 응답하여 수행된다.

[0109] 도 13 은 디바이스 (102) 가 카메라 디바이스 (1302) 에 대응하는 휴대용 전자 디바이스를 포함하는 구현 (1300) 을 도시한다. 스피치 생성기 (120), 마이크로폰 (182), 또는 그 조합은 카메라 디바이스 (1302) 에 포함된다. 동작 동안, 구두 커맨드를 수신하고 스피치 생성기 (120) 의 동작을 통해 합성된 스피치를 생성하는 것 (또는 도 6 의 스피치 인식 네트워크 (672) 의 동작을 통해 스피치를 인식하는 것) 에 응답하여, 카메라 디바이스 (1302) 는 예시적인 예들로서, 이미지 또는 비디오 캡처 설정들, 이미지 또는 비디오 재생 설정들, 또는 이미지 또는 비디오 캡처 명령들을 조정하는 것과 같은, 스피킹된 사용자 커맨드들에 응답하여 동작들을 실행할 수 있다.

[0110] 도 14 는 디바이스 (102) 가 가상 현실, 증강 현실, 또는 혼합 현실 헤드셋 (1402) 에 대응하는 휴대용 전자 디바이스를 포함하는 구현 (1400) 을 도시한다. 스피치 생성기 (120), 마이크로폰 (182), 또는 이들의 조합은 헤드셋 (1402) 에 통합된다. 특정한 양태에서, 헤드셋 (1402) 은 주로 사용자의 스피치를 캡처하도록 포지셔닝된 마이크로폰 (182) 을 포함한다. 합성된 스피치 생성은 헤드셋 (1402) 의 마이크로폰 (182) 으로부터 수신된 오디오 신호들에 기초하여 수행될 수 있다. 시각적 인터페이스 디바이스는 헤드셋 (1402) 이 착용되는 동안 사용자에게 증강 현실 또는 가상 현실 이미지들 또는 장면들의 디스플레이를 가능하게 하기 위해 사용자의 눈 앞에 포지셔닝된다. 특정한 예에서, 시각적 인터페이스 디바이스는 오디오 신호에서 검출된 사용자

스피치를 표시하는 통지를 디스플레이하도록 구성된다.

[0111] 도 15 는 디바이스 (102) 가 유인 또는 무인 항공기 디바이스 (예를 들어, 패키지 전달 드론) 로서 예시된 차량 (1502) 에 대응하거나 그 내에 통합되는 구현 (1500) 을 도시한다. 스피치 생성기 (120), 마이크로폰 (182), 또는 이들의 조합은 차량 (1502) 에 통합된다. 합성된 스피치 생성은 차량 (1502) 의 인증된 사용자로부터의 전달 명령들을 위해서와 같은, 차량 (1502) 의 마이크로폰 (182) 으로부터 수신된 오디오 신호들에 기초하여 수행될 수 있다.

[0112] 도 16 은 디바이스 (102) 가 자동차로서 예시된 차량 (1602) 에 대응하거나 그 내에 통합되는 다른 구현 (1600) 을 도시한다. 차량 (1602) 은 스피치 생성기 (120) 를 포함하는 프로세서 (190) 를 포함한다. 차량 (1602) 은 또한 마이크로폰 (182) 을 포함한다. 마이크로폰 (182) 은 차량 (1602) 의 조작자의 발화들을 캡처하도록 포지셔닝된다. 합성된 스피치 생성은 차량 (1602) 의 마이크로폰 (182) 으로부터 수신된 오디오 신호들에 기초하여 수행될 수 있다. 일부 구현들에서, 합성된 스피치 생성은 인증된 승객으로부터의 음성 커맨드에 대해서와 같은, 내부 마이크로폰들 (예를 들어, 마이크로폰 (182)) 로부터 수신된 오디오 신호에 기초하여 수행될 수 있다. 예를 들어, 합성된 스피치 생성은 차량 (1602) 의 조작자로부터의 음성 커맨드를 검출하는데 (예를 들어, 볼륨을 5 로 설정하거나 자율-주행 차량에 대한 목적지를 설정하는데) 사용될 수 있다. 일부 구현들에서, 합성된 스피치 생성은 차량의 인증된 사용자와 같은 외부 마이크로폰들 (예를 들어, 마이크로폰 (182)) 로부터 수신된 오디오 신호에 기초하여 수행될 수 있다. 특정한 구현에서, 구두 커맨드를 수신하고 스피치 생성기 (120) 의 동작을 통해 합성된 스피치를 생성하는 것에 응답하여, 음성 활성화 시스템 (162) 은, 예컨대 디스플레이 (1620) 또는 하나 이상의 스피커들 (예를 들어, 스피커 (1610)) 을 통해 피드백 또는 정보를 제공함으로써, 합성된 스피치 신호 (131) 에서 검출된 하나 이상의 키워드들 (예를 들어, "잠금 해제", "엔진 시동", "뮤직 재생", "일기 예보 디스플레이", 또는 다른 음성 커맨드) 에 기초하여 차량 (1602) 의 하나 이상의 동작들을 개시한다.

[0113] 도 17 을 참조하면, 합성된 스피치 생성의 방법 (1700) 의 특정한 구현이 나타나 있다. 특정한 양태에서, 방법 (1700) 의 하나 이상의 동작들은 스타일 데이터 생성기 (126), 멀티-인코더 (192), 디코더 (194), 멀티-인코더 변환기 (128), 스피치 생성기 (120), 하나 이상의 프로세서들 (190), 디바이스 (102), 도 1 의 시스템 (100), 사람 인코더 (206), 감정 인코더 (208), 스피치 레이트 인코더 (210), 인코더 프리-네트 (212), 도 4a 의 사람 인코더 (420), 감정 인코더 (430), 레이트 인코더 (440), 입력 표현 인코더 (450), 특성 인코더 (451), 도 5 의 인코더 (500), 또는 이들의 조합 중 적어도 하나에 의해 수행된다.

[0114] 방법 (1700) 은 1702 에서, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 것을 포함한다. 예를 들어, 도 1 의 스타일 데이터 생성기 (126) 는 도 1 을 참조하여 설명된 바와 같이, 타겟 스피치 특성들 (141) 을 표시하는 하나 이상의 제어 파라미터들 (125) 을 수신한다.

[0115] 방법 (1700) 은 또한, 1704 에서, 멀티-인코더를 사용하여, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을 프로세싱하는 것을 포함한다. 예를 들어, 멀티-인코더 (192) 는 하나 이상의 제어 파라미터들 (125) 에 기초하여 입력 표현 (123) 을 프로세싱하여, 도 1 을 참조하여 설명된 바와 같이, 타겟 스피치 특성들 (141) 에 기초하여 스피치의 버전을 표현하는 합성된 스피치 신호 (131) 에 대응하는 인코딩된 데이터 (193) 를 생성한다.

[0116] 방법 (1700) 은 동일한 입력 신호 (121) 에 기초하여 다양한 사람들의 스피치 특성들, 스피치의 다양한 레이트들 및 다양한 감정들에 대응하는 다중의 합성된 스피치 신호들의 생성을 가능하게 한다. 비교적 적은 스피킹된 워드들을 사용하여 크고 다양한 합성된 스피치 데이터세트가 생성될 수 있다.

[0117] 도 17 의 방법 (1700) 은 필드 프로그램가능 게이트 어레이 (FPGA) 디바이스, 주문형 집적 회로 (ASIC), 프로세싱 유닛, 예컨대 중앙 프로세싱 유닛 (CPU), DSP, 제어기, 다른 하드웨어 디바이스, 펌웨어 디바이스, 또는 이들의 임의의 조합에 의해 구현될 수도 있다. 일 예로서, 도 17 의 방법 (1700) 은 도 19 를 참조하여 설명된 바와 같은, 명령들을 실행하는 프로세서에 의해 수행될 수도 있다.

[0118] 도 18 을 참조하면, 합성된 스피치 생성의 방법 (1800) 의 특정한 구현이 나타나 있다. 특정한 양태에서, 방법 (1800) 의 하나 이상의 동작들은 스타일 데이터 생성기 (126), 멀티-인코더 (192), 디코더 (194), 멀티-인코더 변환기 (128), 스피치 생성기 (120), 스피치 합성기 (130), 하나 이상의 프로세서들 (190), 디바이스 (102), 도 1 의 시스템 (100), 사람 인코더 (206), 감정 인코더 (208), 스피치 레이트 인코더 (210), 인코더

프리-네트 (212), 도 4a 의 사람 인코더 (420), 감정 인코더 (430), 레이트 인코더 (440), 입력 표현 인코더 (450), 특성 인코더 (451), 도 5 의 인코더 (500), 또는 이들의 조합 중 적어도 하나에 의해 수행된다.

- [0119] 방법 (1800) 은 1802 에서, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 것을 포함한다. 예를 들어, 도 1 의 스타일 데이터 생성기 (126) 는 도 1 을 참조하여 설명된 바와 같이, 타겟 스피치 특성들 (141) 을 표시하는 하나 이상의 제어 파라미터들 (125) 을 수신한다.
- [0120] 방법 (1800) 은 또한 1804 에서, 하나 이상의 제어 파라미터들에 기초하여 병합된 스타일 데이터를 생성하는 것을 포함한다. 예를 들어, 도 1 의 스타일 데이터 생성기 (126) 는 도 1 을 참조하여 설명된 바와 같이, 하나 이상의 제어 파라미터들 (125) 에 기초하여 병합된 스타일 데이터 (127) 를 생성한다.
- [0121] 방법 (1800) 은, 1806 에서, 멀티-인코더의 제 1 인코더에서, 제 1 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들과 독립적으로 입력 표현을 인코딩하는 것을 더 포함한다. 예를 들어, 멀티-인코더 (192) 의 입력 표현 인코더 (450) 는 도 4a 를 참조하여 설명된 바와 같이, 제 1 인코딩된 데이터 (458) 를 생성하기 위해 하나 이상의 제어 파라미터들 (125) 과는 독립적으로 입력 표현 (123) 을 인코딩한다.
- [0122] 방법 (1800) 은 또한, 1808 에서, 멀티-인코더의 하나 이상의 제 2 인코더들에서, 제 2 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 입력 표현을 인코딩하는 것을 포함한다. 예를 들어, 도 4a 의 사람 인코더 (420), 감정 인코더 (430), 및 레이트 인코더 (440) 는 도 4a 를 참조하여 설명된 바와 같이, 제 2 인코딩된 데이터 (예를 들어, 사람 인코딩된 데이터 (428), 감정 인코딩된 데이터 (438), 및 스피치 레이트 인코딩된 데이터 (448)) 를 생성하기 위해 하나 이상의 제어 파라미터들 (125) 에 기초하여 입력 표현 (123) 을 인코딩한다. 다른 예로서, 도 4b 의 특성 인코더 (451) 는 도 4b 를 참조하여 설명된 바와 같이, 제 2 인코딩된 데이터 (459) 를 생성하기 위해 하나 이상의 제어 파라미터들 (125) 에 기초하여 입력 표현 (123) 을 인코딩한다.
- [0123] 방법 (1800) 은, 1810 에서, 디코더의 디코더 네트워크에서, 제 1 인코딩된 데이터 및 제 2 인코딩된 데이터에 기초하여 출력 스펙트럼 데이터를 생성하는 것을 더 포함한다. 예를 들어, 디코더 (194)(예를 들어, 도 4a 의 멀티-헤드 어텐션 (478), 멀티-헤드 어텐션 (480), 멀티-헤드 어텐션 (482), 및 멀티-헤드 어텐션 (484) 그리고 도 4b 의 멀티-헤드 어텐션 (485) 과 같은 복수의 디코더 네트워크들을 포함함) 는 도 1 및 도 4a 및 도 4b 를 참조하여 설명된 바와 같이, 인코딩된 데이터 (193) 에 기초하여 출력 스펙트럼 데이터 (129) 를 생성한다.
- [0124] 방법 (1800) 은 또한, 1812 에서, 스피치 합성기에서 그리고 출력 스펙트럼 데이터에 기초하여, 타겟 스피치 특성들에 기초한 스피치의 버전을 표현하는 오디오 신호를 생성하는 것을 포함한다. 예를 들어, 도 1 의 스피치 합성기 (130) 는, 도 1 을 참조하여 설명된 바와 같이, 출력 스펙트럼 데이터 (129) 에 기초하여, 타겟 스피치 특성들 (141) 에 기초하여 스피치의 버전을 표현하는 합성된 스피치 신호 (131) 를 생성한다.
- [0125] 방법 (1800) 은 동일한 입력 신호 (121) 에 기초하여 다양한 사람들의 스피치 특성들, 스피치의 다양한 레이트들 및 다양한 감정들에 대응하는 다중의 합성된 스피치 신호들의 생성을 가능하게 한다. 비교적 적은 스피킹된 워드들을 사용하여 크고 다양한 합성된 스피치 데이터세트가 생성될 수 있다.
- [0126] 도 18 의 방법 (1800) 은 필드 프로그램가능 게이트 어레이 (FPGA) 디바이스, 주문형 집적 회로 (ASIC), 프로세싱 유닛, 예컨대 중앙 프로세싱 유닛 (CPU), DSP, 제어기, 다른 하드웨어 디바이스, 펌웨어 디바이스, 또는 이들의 임의의 조합에 의해 구현될 수도 있다. 일 예로서, 도 18 의 방법 (1800) 은 도 19 를 참조하여 설명된 바와 같은, 명령들을 실행하는 프로세서에 의해 수행될 수도 있다.
- [0127] 도 19 를 참조하면, 디바이스의 특정한 예시적인 구현의 블록 다이어그램이 도시되고 일반적으로 1900 으로 지정된다. 다양한 구현들에서, 디바이스 (1900) 는 도 19 에 예시된 것보다 더 많거나 또는 더 적은 컴포넌트들을 가질 수도 있다. 예시적인 구현에서, 디바이스 (1900) 는 디바이스 (102) 에 대응할 수도 있다. 예시적인 구현에서, 디바이스 (1900) 는 도 1 내지 도 18 을 참조하여 설명된 하나 이상의 동작들을 수행할 수도 있다.
- [0128] 특정한 구현에서, 디바이스 (1900) 는 프로세서 (1906)(예를 들어, 중앙 프로세싱 유닛 (CPU)) 를 포함한다. 디바이스 (1900) 는 하나 이상의 부가적인 프로세서 (1910) (예를 들어, 하나 이상의 DSP) 를 포함할 수도 있다. 특정한 양태에서, 도 1 의 하나 이상의 프로세서들 (190) 은 프로세서 (1906), 프로세서들 (1910), 또는 이들의 조합에 대응한다. 프로세서들 (1910) 은 음성 코더 ("보코더") 인코더 (1936), 보코더 디코더 (1938), 스피치 생성기 (120), 네트워크 훈련기 (670), 또는 이들의 조합을 포함하는 스피치 및 뮤직 코더-디코

더 (코덱)(CODEC)(1908)를 포함할 수도 있다.

- [0129] 디바이스 (1900)는 메모리 (132) 및 코덱 (1934)을 포함할 수도 있다. 메모리 (132)는 스피치 생성기 (120), 네트워크 훈련기 (670), 또는 양자 모두를 참조하여 설명된 기능성을 구현하기 위해 하나 이상의 부가 프로세서들 (1910)(또는 프로세서 (1906))에 의해 실행가능한 명령들 (1956)을 포함할 수도 있다. 디바이스 (1900)는 트랜시버 (1950)를 통해, 안테나 (1952)에 커플링된 모뎀 (1970)을 포함할 수도 있다.
- [0130] 디바이스 (1900)는 디스플레이 제어기 (1926)에 커플링된 디스플레이 (1928)를 포함할 수도 있다. 스피커 (1992), 마이크로폰 (182), 또는 양자 모두는 코덱 (1934)에 커플링될 수도 있다. 코덱 (1934)은 디지털-투-아날로그 컨버터 (DAC)(1902), 아날로그-투-디지털 컨버터 (ADC)(1904), 또는 양자 모두를 포함할 수도 있다. 특정한 구현에서, 코덱 (1934)은 마이크로폰 (182)으로부터 아날로그 신호들을 수신하고, 아날로그 신호들을 아날로그-디지털 변환기 (1904)를 사용하여 디지털 신호들로 변환하고, 디지털 신호들을 스피치 및 뮤직 코덱 (1908)에 제공할 수도 있다. 스피치 및 뮤직 코덱 (1908)은 디지털 신호들을 프로세싱할 수도 있고, 디지털 신호들은 스피치 생성기 (120)에 의해 추가로 프로세싱될 수도 있다. 특정한 구현에서, 스피치 및 뮤직 코덱 (1908)은 코덱 (1934)에 디지털 신호들을 제공할 수도 있다. 코덱 (1934)은 디지털-투-아날로그 변환기 (1902)를 사용하여 디지털 신호들을 아날로그 신호들로 변환할 수도 있고 아날로그 신호들을 스피커들 (1992)에 제공할 수도 있다.
- [0131] 특정한 구현에서, 디바이스 (1900)는 시스템-인-패키지 또는 시스템-온-칩 디바이스 (1922)에 포함될 수도 있다. 특정한 구현에서, 메모리 (132), 프로세서 (1906), 프로세서들 (1910), 디스플레이 제어기 (1926), 코덱 (1934), 및 모뎀 (1970)은 시스템-인-패키지 또는 시스템-온-칩 디바이스 (1922)에 포함된다. 특정한 구현에서, 입력 디바이스 (1930) 및 전력 공급장치 (1944)는 시스템-온-칩 디바이스 (1922)에 커플링된다. 게다가, 특정한 구현에서, 도 19에 도시된 바와 같이, 디스플레이 (1928), 입력 디바이스 (1930), 스피커 (1992), 마이크로폰 (182), 안테나 (1952), 및 전력 공급장치 (1944)는 시스템-온-칩 디바이스 (1922)외부에 있다. 특정한 구현에서, 디스플레이 (1928), 입력 디바이스 (1930), 스피커 (1992), 마이크로폰 (182), 안테나 (1952), 및 전력 공급기 (1944)의 각각은, 인터페이스 또는 제어기와 같은 시스템-온-칩 디바이스 (1922)의 컴포넌트에 커플링될 수도 있다.
- [0132] 디바이스 (1900)는 가상 어시스턴트, 홈 어플라이언스, 스마트 디바이스, 사물 인터넷 (IoT) 디바이스, 통신 디바이스, 헤드셋, 차량, 컴퓨터, 디스플레이 디바이스, 텔레비전, 게이밍 콘솔, 뮤직 플레이어, 라디오, 비디오 플레이어, 엔터테인먼트 유닛, 개인용 미디어 플레이어, 디지털 비디오 플레이어, 카메라, 내비게이션 디바이스, 스마트 스피커, 스피커 바, 모바일 통신 디바이스, 스마트폰, 셀룰러 폰, 랩톱 컴퓨터, 태블릿, 개인용 디지털 보조기, 디지털 비디오 디스크 (DVD) 플레이어, 튜너, 증강 현실 헤드셋, 가상 현실 헤드셋, 항공 차량, 홈 자동화 시스템, 음성-활성화 디바이스, 무선 스피커 및 음성 활성화 디바이스, 휴대용 전자 디바이스, 자동차, 차량, 컴퓨팅 디바이스, 가상 현실 (VR) 디바이스, 기지국, 모바일 디바이스, 또는 이들의 임의의 조합을 포함할 수도 있다.
- [0133] 설명된 구현들과 함께, 장치는 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 수단을 포함한다. 예를 들어, 수신하는 수단은 도 1의 스타일 데이터 생성기 (126), 스피치 생성기 (120), 하나 이상의 프로세서들 (190), 디바이스 (102), 트랜시버 (1950), 모뎀 (1970), 입력 디바이스 (1930), 디바이스 (1900)의 입력 인터페이스, 하나 이상의 제어 파라미터들을 수신하도록 구성된 하나 이상의 다른 회로들 또는 컴포넌트들, 또는 이들의 임의의 조합을 포함한다.
- [0134] 장치는 또한, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 수단을 포함한다. 예를 들어, 프로세싱하는 수단은 도 1의 멀티-인코더 (192), 멀티-인코더 변환기 (128), 스피치 생성기 (120), 하나 이상의 프로세서들 (190), 디바이스 (102), 스피치 및 뮤직 코덱 (1908), 프로세서 (1910), 프로세서 (1906), 디바이스 (1900), 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을 프로세싱하도록 구성된 하나 이상의 다른 회로들 또는 컴포넌트들, 또는 이들의 임의의 조합을 포함한다.
- [0135] 일부 구현들에서, 비밀시적 컴퓨터 관독가능 매체 (예를 들어, 메모리 (132)와 같은 컴퓨터 관독가능 저장 디바이스)는, 하나 이상의 프로세서들 (예를 들어, 하나 이상의 프로세서들 (1910) 또는 프로세서 (1906))에 의해 실행될 때, 하나 이상의 프로세서들로 하여금, 타겟 스피치 특성들 (예를 들어, 타겟 스피치 특성들 (141))을 표시하는 하나 이상의 제어 파라미터들 (예를 들어, 제어 파라미터들 (125))을 수신하게 하는 명령들 (예를

들어, 명령들 (1956)) 을 포함한다. 명령들은, 하나 이상의 프로세서들에 의해 실행될 때, 또한 하나 이상의 프로세서들로 하여금, 멀티-인코더 (예를 들어, 멀티-인코더 (192)) 를 사용하여, 타겟 스피치 특성들 (예를 들어, 타겟 스피치 특성들 (141)) 에 기초하여 스피치의 버전을 표현하는 오디오 신호 (예를 들어, 합성된 스피치 신호 (131)) 에 대응하는 인코딩된 데이터 (예를 들어, 인코딩된 데이터 (193)) 를 생성하기 위해 하나 이상의 제어 파라미터들 (예를 들어, 하나 이상의 제어 파라미터들 (125)) 에 기초하여 스피치의 입력 표현 (예를 들어, 입력 표현 (123)) 을 프로세싱하게 한다.

[0136] 본 개시의 특정한 양태들은 상호관련된 조항들의 제 1 세트로 하기에서 설명된다:

[0137] 조항 1 에 따르면, 스피치 생성을 위한 디바이스는 하나 이상의 프로세서들을 포함하고, 하나 이상의 프로세서들은 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하고; 그리고 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하도록 구성된다.

[0138] 조항 2 는 조항 1 의 디바이스를 포함하고, 제어 파라미터들은, 스피치 특성들이 사용될 타겟 사람, 타겟 감정, 타겟 스피치 레이트, 또는 이들의 조합을 표시한다.

[0139] 조항 3 은 조항 1 또는 조항 2 의 디바이스를 포함하고, 하나 이상의 프로세서들은 추가로, 상기 하나 이상의 제어 파라미터들에 기초하여 병합된 스타일 데이터를 생성하도록 구성되고, 병합된 스타일 데이터는 입력 표현의 프로세싱 동안 멀티-인코더에 의해 사용된다.

[0140] 조항 4 는 조항들 1 내지 3 중 임의의 것의 디바이스를 포함하고, 멀티-인코더는, 제 1 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들과는 독립적으로 입력 표현을 인코딩하도록 구성된 제 1 인코더; 및 제 2 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 입력 표현을 인코딩하도록 구성된 하나 이상의 제 2 인코더들을 포함하며, 인코딩된 데이터는 제 1 인코딩된 데이터 및 제 2 인코딩된 데이터를 포함한다.

[0141] 조항 5 는 조항 4 의 디바이스를 포함하고, 하나 이상의 프로세서들은 추가로, 스피치 특성 인코더에서, 인코딩된 입력 스피치 표현을 생성하기 위해 하나 이상의 제어 파라미터들 중 적어도 하나에 기초하여 입력 표현을 프로세싱하고; 인코더 프리-네트워크에서, 인코딩된 입력 스피치 표현에 적어도 부분적으로 기초하여 병합된 스타일 데이터를 생성하고; 제 1 인코딩된 데이터를 생성하기 위해 제 1 인코더에 입력 표현을 제공하며; 그리고 타겟 제 2 인코딩된 데이터를 생성하기 위해 하나 이상의 제 2 인코더들에 병합된 스타일 데이터를 제공하도록 구성된다.

[0142] 조항 6 은 조항 4 또는 조항 5 의 디바이스를 포함하고, 멀티-인코더 및 디코더를 포함하는 멀티-인코더 변환기를 더 포함하며, 제 1 인코더는 제 1 어텐션 네트워크를 포함하고, 하나 이상의 제 2 인코더들의 각각은 제 2 어텐션 네트워크를 포함하며, 디코더는 디코더 어텐션 네트워크를 포함한다.

[0143] 조항 7 은 조항 6 의 디바이스를 포함하고, 제 1 인코더는, 제 1 어텐션 네트워크를 포함하는 제 1 계층으로서, 제 1 어텐션 네트워크는 제 1 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 1 어텐션 네트워크를 포함하는 제 1 계층; 및 제 1 뉴럴 네트워크를 포함하는 제 2 계층을 포함하고, 하나 이상의 제 2 인코더들의 각각은, 제 2 어텐션 네트워크를 포함하는 제 1 계층으로서, 제 2 어텐션 네트워크는 제 2 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 2 어텐션 네트워크를 포함하는 제 1 계층; 및 제 2 뉴럴 네트워크를 포함하는 제 2 계층을 포함한다.

[0144] 조항 8 은 조항들 4 내지 7 중 임의의 것이 디바이스를 포함하고, 멀티-인코더에 커플링된 디코더로서, 디코더는 제 1 인코딩된 데이터 및 타겟 제 2 인코딩된 데이터에 기초하여 출력 스펙트럼 데이터를 생성하도록 구성되는 디코더 네트워크를 포함하는, 상기 디코더; 및 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호를, 출력 스펙트럼 데이터에 기초하여, 생성하도록 구성된 스피치 합성기를 더 포함한다.

[0145] 조항 9 는 조항 8 의 디바이스를 포함하고, 디코더 네트워크는 디코더 어텐션 네트워크를 포함하고, 디코더 어텐션 네트워크는, 제 1 인코딩된 데이터를 프로세싱하도록 구성된 제 1 멀티-헤드 어텐션 네트워크; 타겟 제 2 인코딩된 데이터를 프로세싱하도록 구성된 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들; 및 제 1 멀티-헤드 어텐션 네트워크 및 상기 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들의 출력들을 결합하도록 구성된 결합기를 포함한다.

[0146] 조항 10 은 조항 9 의 디바이스를 포함하고, 디코더는, 디코더 어텐션 네트워크의 입력에 커플링된 마스킹된 멀

터-헤드 어텐션 네트워크; 및 디코더 어텐션 네트워크의 출력에 커플링된 디코더 뉴럴 네트워크를 더 포함한다.

- [0147] 조항 11 은 조항들 1 내지 10 중 임의의 것의 디바이스를 포함하고, 하나 이상의 프로세서들은 추가로, 오디오 신호로부터 하나 이상의 추정된 제어 파라미터들을 생성하고; 그리고 하나 이상의 제어 파라미터들과 하나 이상의 추정된 제어 파라미터들의 비교에 기초하여, 멀티-인코더, 하나 이상의 스피치 특성 인코더들, 인코더 프리-네트워크, 디코더 네트워크, 또는 이들의 조합의 하나 이상의 뉴럴 네트워크 가중치들을 훈련하도록 구성된다.
- [0148] 조항 12 는 조항들 1 내지 11 중 임의의 것의 디바이스를 포함하고, 하나 이상의 프로세서들은 추가로, 입력 스피치 신호를 수신하고; 그리고 입력 스피치 신호에 기초하여 입력 표현을 생성하도록 구성된다.
- [0149] 조항 13 은 조항들 1 내지 11 중 임의의 것의 디바이스를 포함하고, 하나 이상의 프로세서들은 추가로, 입력 표현을 수신하도록 구성된다.
- [0150] 조항 14 는 조항들 1 내지 13 중 임의의 것의 디바이스를 포함하고, 입력 표현은 텍스트, 멜-스케일(mel-scale) 스펙트로그램들, 기본 주파수 (F0) 특징들, 또는 이들의 조합을 포함한다.
- [0151] 본 개시의 특정한 양태들은 상호관련된 조항들의 제 2 세트로 하기에서 설명된다:
- [0152] 조항 15 에 따르면, 스피치 생성 방법으로서, 디바이스에서, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 단계; 및 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 상기 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 단계를 포함한다.
- [0153] 조항 16 은 조항 15 의 방법을 포함하고, 제어 파라미터들은, 스피치 특성들이 사용될 타겟 사람, 타겟 감정, 타겟 스피치 레이트, 또는 이들의 조합을 표시한다.
- [0154] 조항 17 은 조항 15 또는 조항 16 의 방법을 포함하고, 디바이스에서, 상기 하나 이상의 제어 파라미터들에 기초하여 병합된 스타일 데이터를 생성하는 단계를 더 포함하며, 병합된 스타일 데이터는 입력 표현의 프로세싱 동안 멀티-인코더에 의해 사용된다.
- [0155] 조항 18 은 조항들 15 내지 17 중 임의의 것의 방법을 포함하고, 멀티-인코더의 제 1 인코더에서, 제 1 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들과 독립적으로 입력 표현을 인코딩하는 단계; 및 멀티-인코더의 하나 이상의 제 2 인코더들에서, 타겟 제 2 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 입력 표현을 인코딩하는 단계를 더 포함하며, 오디오 신호 인코딩된 데이터는 제 1 인코딩된 데이터 및 타겟 제 2 인코딩된 데이터에 기초한다.
- [0156] 조항 19 는 조항 18 의 방법을 포함하고, 스피치 특성 인코더에서, 인코딩된 입력 스피치 표현을 생성하기 위해 하나 이상의 제어 파라미터들 중 적어도 하나에 기초하여 입력 표현을 프로세싱하는 단계; 인코더 프리-네트워크에서, 인코딩된 입력 스피치 표현에 적어도 부분적으로 기초하여 병합된 스타일 데이터를 생성하는 단계; 제 1 인코딩된 데이터를 생성하기 위해 제 1 인코더에 입력 표현을 제공하는 단계; 및 타겟 제 2 인코딩된 데이터를 생성하기 위해 하나 이상의 제 2 인코더들에 병합된 스타일 데이터를 제공하는 단계를 더 포함한다.
- [0157] 조항 20 은 조항 18 의 방법을 포함하고, 멀티-인코더 및 디코더가 멀티-인코더 변환기에 포함되고, 제 1 인코더는 제 1 어텐션 네트워크를 포함하고, 하나 이상의 제 2 인코더들의 각각은 제 2 어텐션 네트워크를 포함하며, 디코더는 디코더 어텐션 네트워크를 포함한다.
- [0158] 조항 21 은 조항 20 의 방법을 포함하고, 제 1 인코더는, 제 1 어텐션 네트워크를 포함하는 제 1 계층으로서, 제 1 어텐션 네트워크는 제 1 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 1 어텐션 네트워크를 포함하는 제 1 계층; 및
- [0159] 제 1 뉴럴 네트워크를 포함하는 제 2 계층을 포함하고, 하나 이상의 제 2 인코더들의 각각은, 제 2 어텐션 네트워크를 포함하는 제 1 계층으로서, 제 2 어텐션 네트워크는 제 2 멀티-헤드 어텐션 네트워크에 대응하는, 상기 제 2 어텐션 네트워크를 포함하는 제 1 계층; 및 제 2 뉴럴 네트워크를 포함하는 제 2 계층을 포함한다.
- [0160] 조항 22 는 조항 18 의 방법을 포함하고, 디코더의 디코더 네트워크에서, 제 1 인코딩된 데이터 및 타겟 제 2 인코딩된 데이터에 기초하여 출력 스펙트럼 데이터를 생성하는 단계; 및 스피치 합성기에서 그리고 출력 스펙트럼 데이터에 기초하여, 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호를 생성하는 단계를 더 포함한다.

- [0161] 조항 23 은 조항 22 의 방법을 포함하고, 디코더 어텐션 네트워크의 제 1 멀티-헤드 어텐션 네트워크에서 제 1 인코딩된 데이터를 프로세싱하는 단계로서, 디코더 네트워크는 디코더 어텐션 네트워크를 포함하는, 상기 제 1 인코딩된 데이터를 프로세싱하는 단계; 디코더 어텐션 네트워크의 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들에서 타겟 제 2 인코딩된 데이터를 프로세싱하는 단계; 및 결합기에서, 제 1 멀티-헤드 어텐션 네트워크 및 하나 이상의 제 2 멀티-헤드 어텐션 네트워크들의 출력들을 결합하는 단계를 더 포함한다.
- [0162] 조항 24 는 조항들 15 내지 23 중 임의의 것의 방법을 포함하고, 디바이스에서, 오디오 신호로부터 하나 이상의 추정된 제어 파라미터들을 생성하는 단계; 및 하나 이상의 제어 파라미터들과 하나 이상의 추정된 제어 파라미터들의 비교에 기초하여, 멀티-인코더, 하나 이상의 스피치 특성 인코더들, 인코더 프리-네트워크, 디코더 네트워크, 또는 이들의 조합의 하나 이상의 뉴럴 네트워크 가중치들을 훈련하는 단계를 더 포함한다.
- [0163] 조항 25 는 조항들 15 내지 24 중 임의의 것의 방법을 포함하고, 디바이스에서 입력 스피치 신호를 수신하는 단계; 및 디바이스에서, 입력 스피치 신호에 기초하여 입력 표현을 생성하는 단계를 더 포함한다.
- [0164] 조항 26 은 조항들 15 내지 24 중 임의의 것의 방법을 포함하고, 디바이스에서 입력 표현을 수신하는 단계를 더 포함한다.
- [0165] 본 개시의 특정한 양태들은 상호관련된 조항들의 제 3 세트로 하기에서 설명된다:
- [0166] 조항 27 에 따르면, 명령들을 저장하는 비일시적 컴퓨터 판독가능 저장 매체로서, 명령들은, 하나 이상의 프로세서들에 의해 실행될 때, 하나 이상의 프로세서들로 하여금, 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하게 하고; 그리고 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하게 한다.
- [0167] 조항 28 은 조항 27 의 비일시적 컴퓨터 판독가능 저장 매체를 포함하고, 입력 표현은 텍스트, 멜-스케일 스펙트로그램들, 기본 주파수 (F0) 특징들, 또는 이들의 조합을 포함한다.
- [0168] 본 개시의 특정한 양태들은 상호관련된 조항들의 제 4 세트로 하기에서 설명된다:
- [0169] 조항 29 에 따르면, 장치는 타겟 스피치 특성들을 표시하는 하나 이상의 제어 파라미터들을 수신하는 수단; 및 타겟 스피치 특성들에 기초하여 스피치의 버전을 표현하는 오디오 신호에 대응하는 인코딩된 데이터를 생성하기 위해 하나 이상의 제어 파라미터들에 기초하여 스피치의 입력 표현을, 멀티-인코더를 사용하여, 프로세싱하는 수단을 포함한다.
- [0170] 조항 30 은 조항 29 의 장치를 포함하고, 수신하는 수단 및 프로세싱하는 수단은 가상 어시스턴트, 홈 어플라이언스, 스마트 디바이스, 사물 인터넷 (IoT) 디바이스, 통신 디바이스, 헤드셋, 차량, 컴퓨터, 디스플레이 디바이스, 텔레비전, 게이밍 콘솔, 뮤직 플레이어, 라디오, 비디오 플레이어, 엔터테인먼트 유닛, 개인용 미디어 플레이어, 디지털 비디오 플레이어, 카메라, 또는 내비게이션 디바이스 중 적어도 하나에 통합된다.
- [0171] 당업자는 추가로, 본 명세서에서 개시된 구현들과 관련하여 설명된 다양한 예시적인 논리 블록들, 구성들, 모듈들, 회로들, 및 알고리즘 단계들이 전자 하드웨어, 프로세서에 의해 실행되는 컴퓨터 소프트웨어, 또는 이들 양자의 조합으로서 구현될 수도 있음을 알 것이다. 다양한 예시적인 컴포넌트들, 블록들, 구성들, 모듈들, 회로들, 및 단계들은 일반적으로 이들의 기능성의 관점에서 위에서 설명되었다. 그러한 기능성이 하드웨어로 구현되는지 또는 프로세서 실행가능 명령들로 구현되는지는 전체 시스템에 부과된 특정한 애플리케이션 및 설계 제약들에 의존한다. 당업자는 설명된 기능성을 각각의 특정한 애플리케이션에 대해 다양한 방식으로 구현할 수도 있지만, 이러한 구현 결정들은 본 개시의 범위로부터 벗어남을 야기하는 것으로 해석되지 않아야 한다.
- [0172] 본 명세서에 개시된 구현들과 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어에서, 프로세서에 의해 실행되는 소프트웨어 모듈에서, 또는 이들 2개의 조합에서 직접 구현될 수도 있다. 소프트웨어 모듈이 랜덤 액세스 메모리 (RAM), 플래시 메모리, 판독 전용 메모리 (ROM), 프로그램가능 판독 전용 메모리 (PROM), 소거가능 프로그램가능 판독 전용 메모리 (EPROM), 전기적 소거가능 프로그램가능 판독 전용 메모리 (EEPROM), 레지스터들, 하드 디스크, 착탈식 디스크, 컴팩트 디스크 판독 전용 메모리 (CD-ROM) 또는 종래에 알려진 임의의 형태의 비일시적 저장 매체에 상주할 수도 있다. 예시적인 저장 매체는 프로세서가 저장 매체로부터 정보를 판독할 수 있고 저장 매체에 정보를 기입할 수도 있도록 프로세서에 커플링된다. 대안으로, 저장 매체는 프로세서에 통합될 수도 있다. 프로세서 및 저장 매체는 주문형 집적 회로 (ASIC) 에 상주할 수도 있다. ASIC 은 컴퓨팅 디바이스 또는 사용자 단말기에 상주할 수도 있다. 대안으로, 프로세서 및 저장 매체는 컴

퓨팅 디바이스 또는 사용자 단말기에서 이산 컴포넌트들로서 상주할 수도 있다.

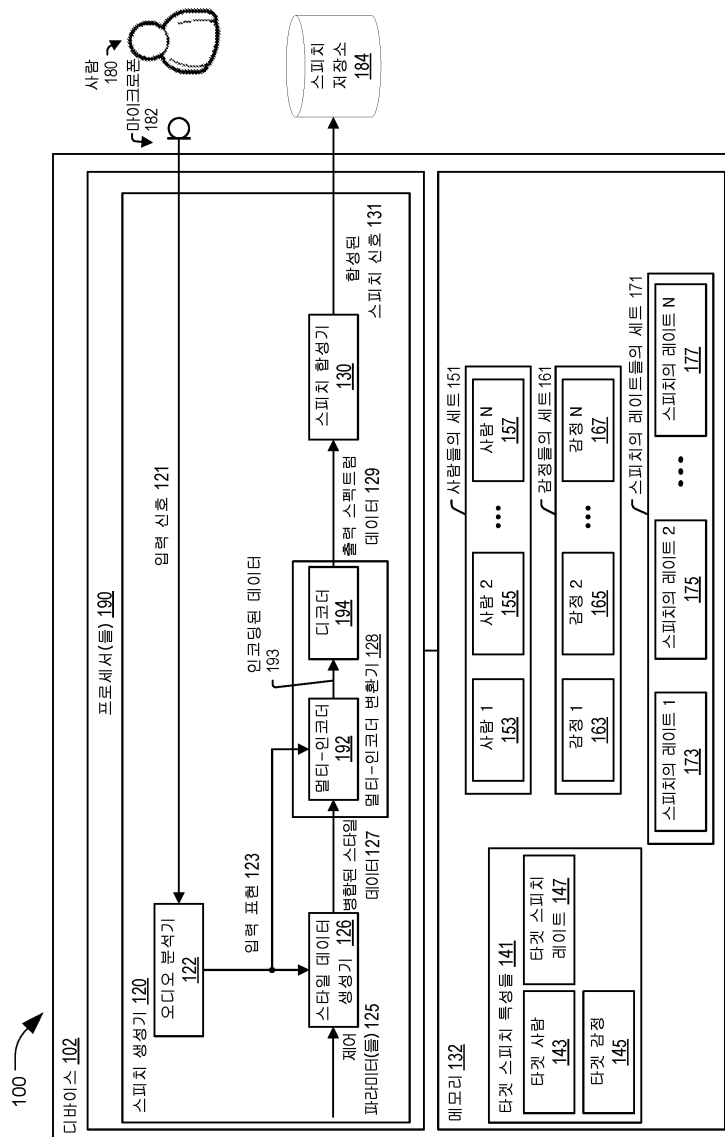
[0173]

개시된 양태들의 이전 설명은 당업자가 개시된 양태들을 제조 또는 사용하는 것을 가능하게 하도록 제공된다.

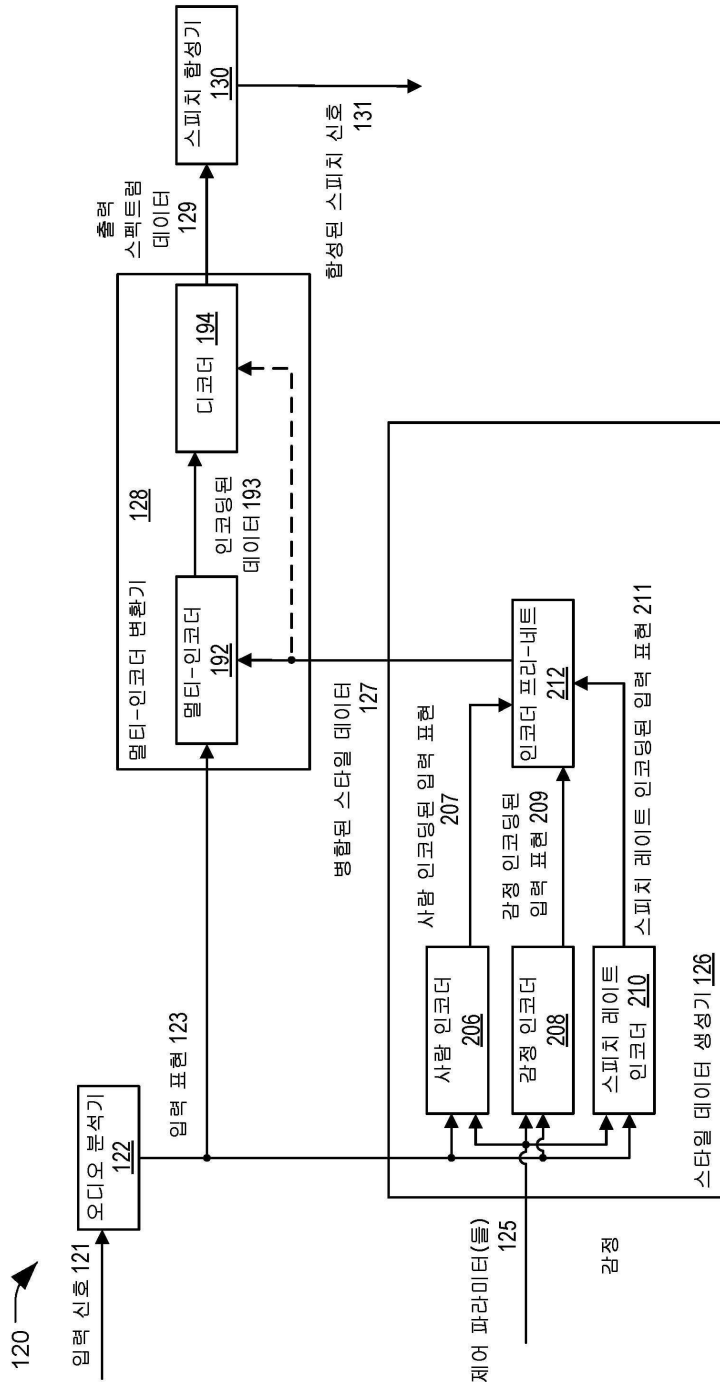
이들 양태들에 대한 다양한 수정들은 당업자에게 용이하게 자명할 것이며, 본 명세서에서 정의된 원리들은 본 개시의 범위로부터 벗어나지 않으면서 다른 양태들에 적용될 수도 있다. 따라서, 본 개시는 본 명세서에 나타난 양태들에 제한되도록 의도되는 것이 아니라, 다음의 청구항들에 의해 정의되는 원리들 및 신규한 특징들에 부합하는 가능한 최광의 범위를 부여하려는 것이다.

도면

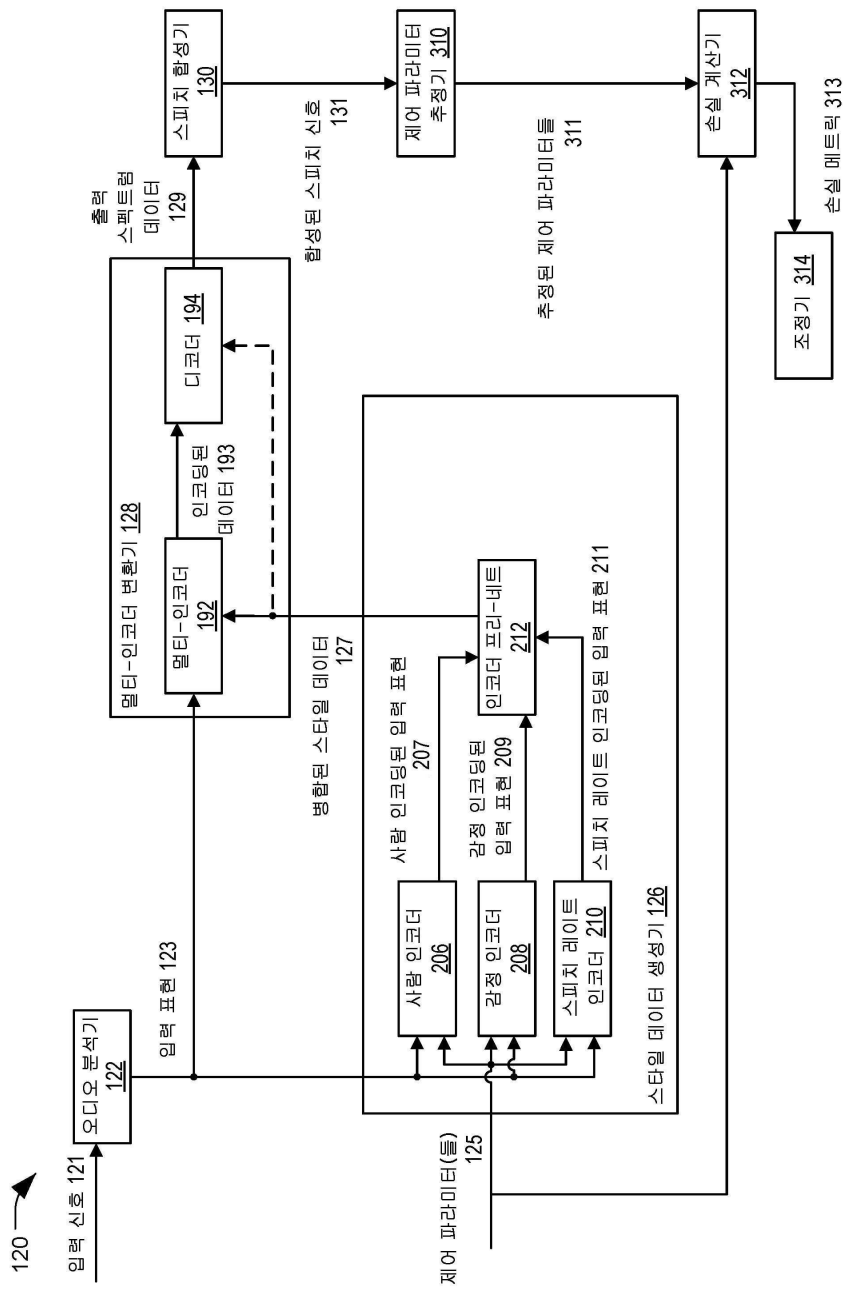
도면1



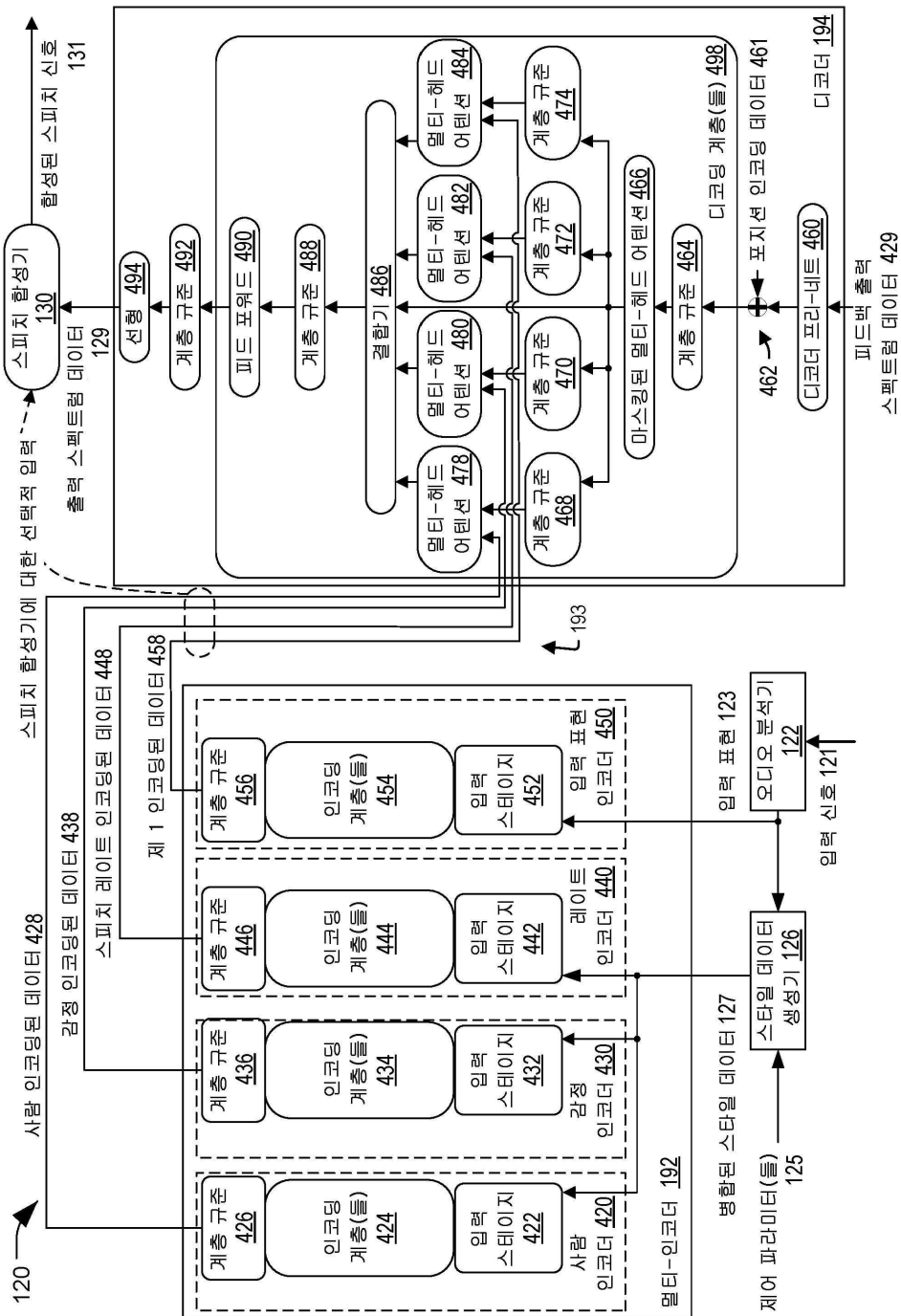
도면2



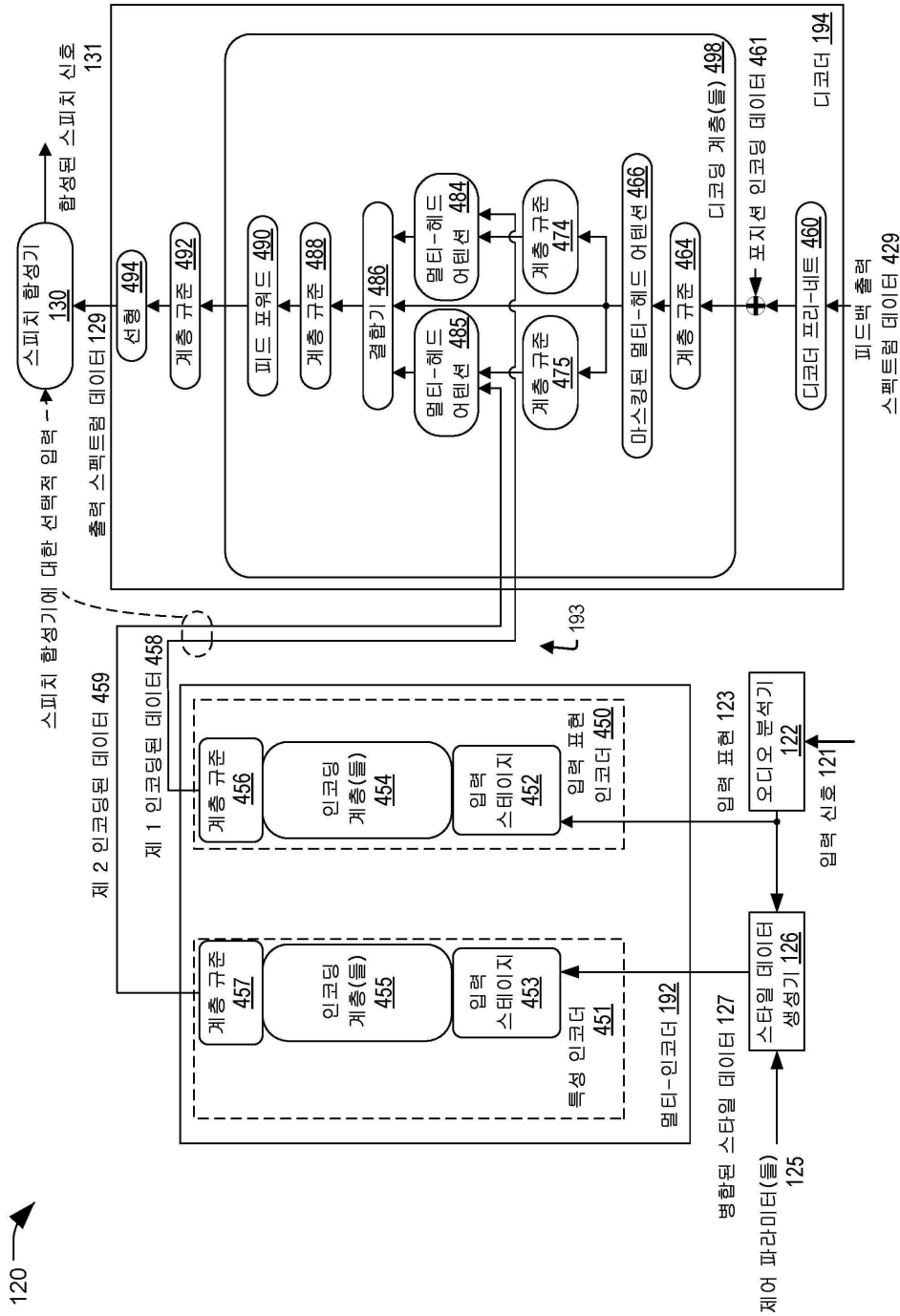
도면3



도면4a

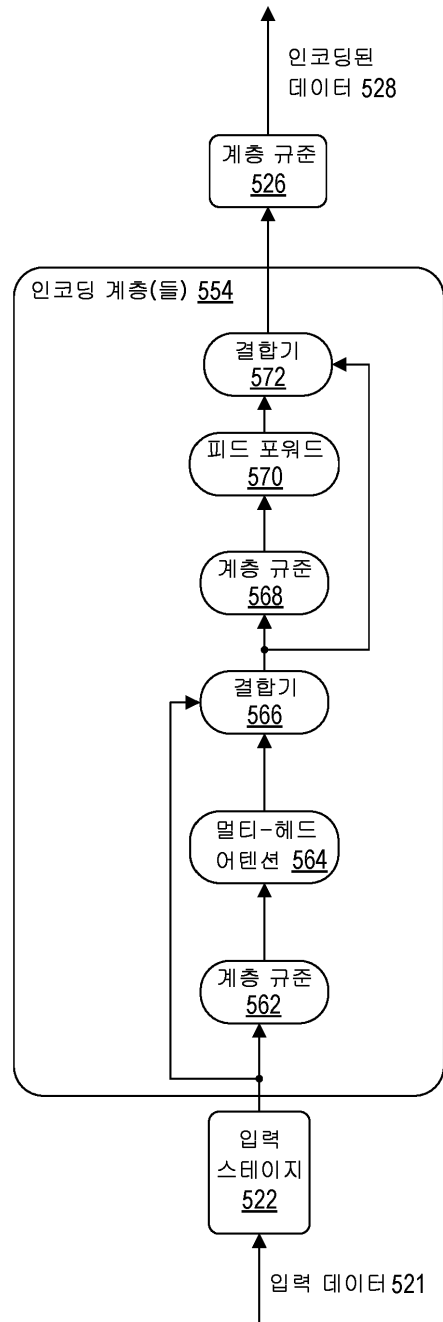


도면4b

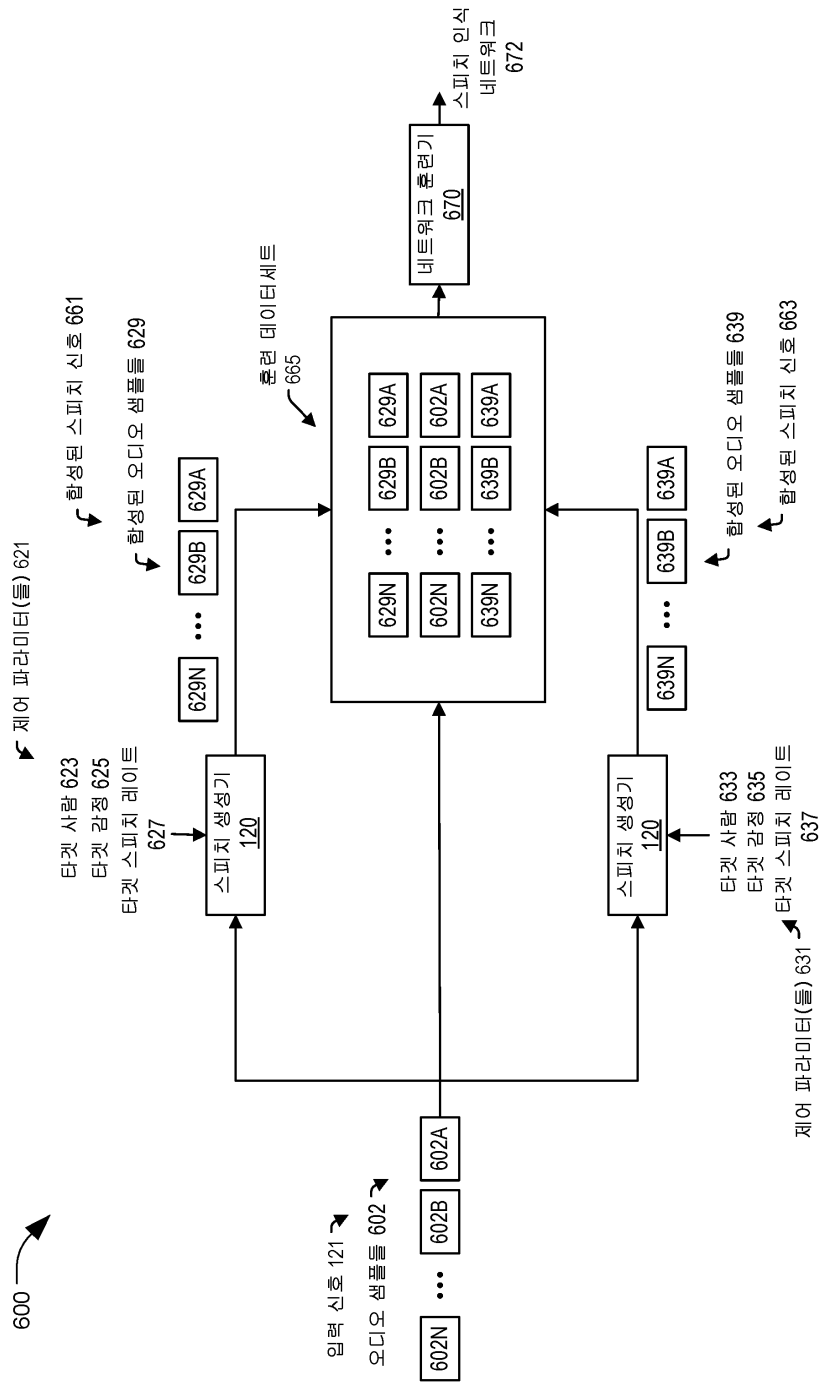


도면5

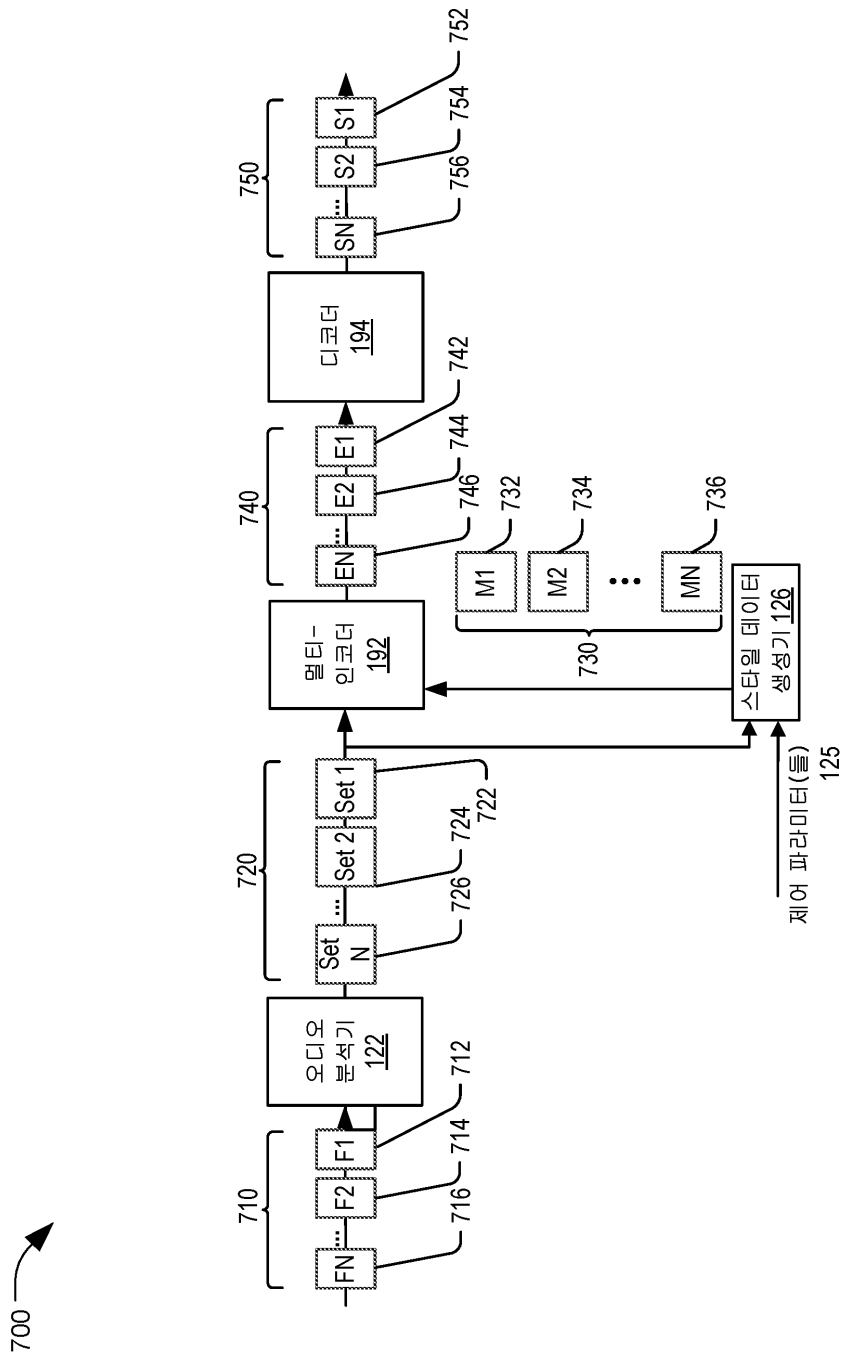
500



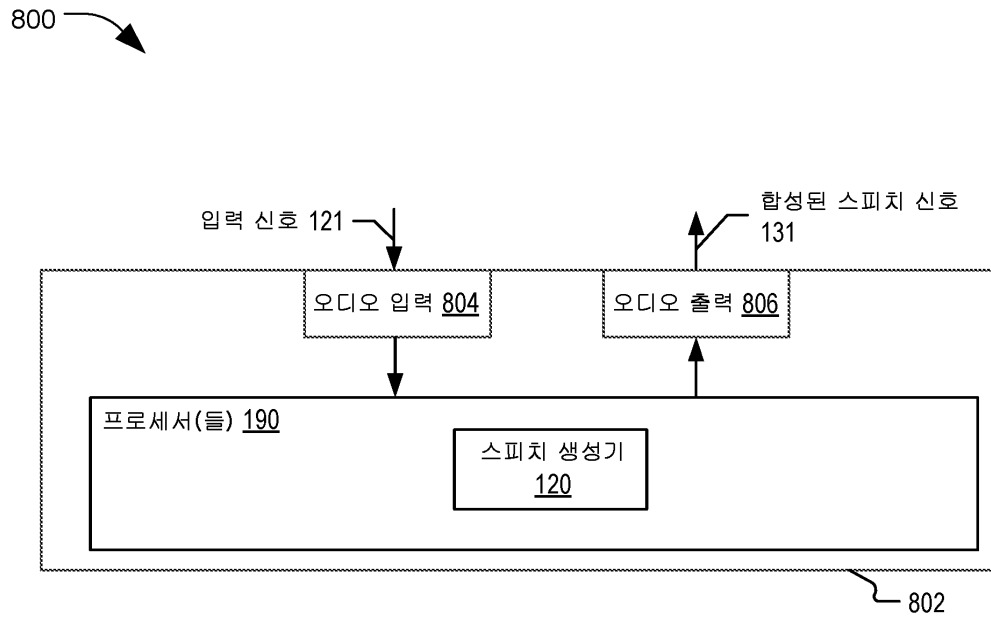
도면6



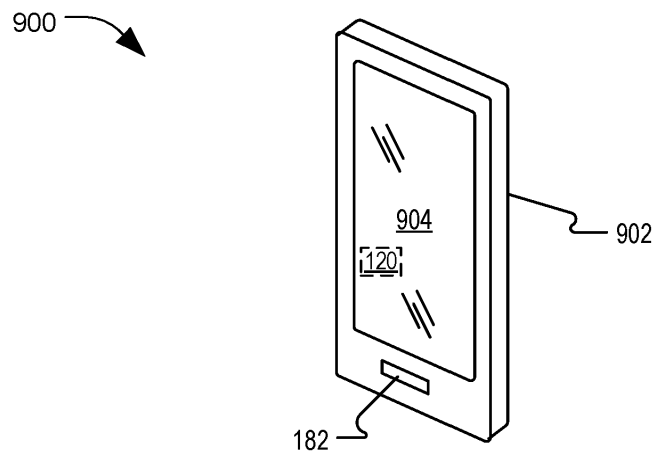
도면7



도면8

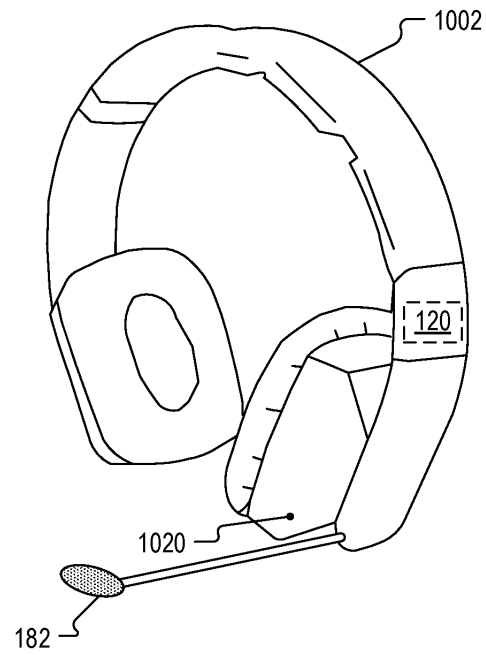


도면9



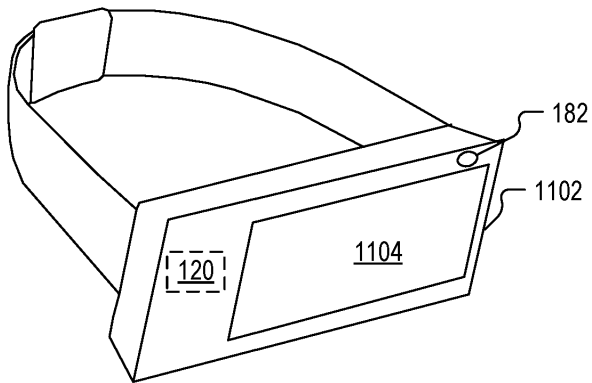
도면10

1000



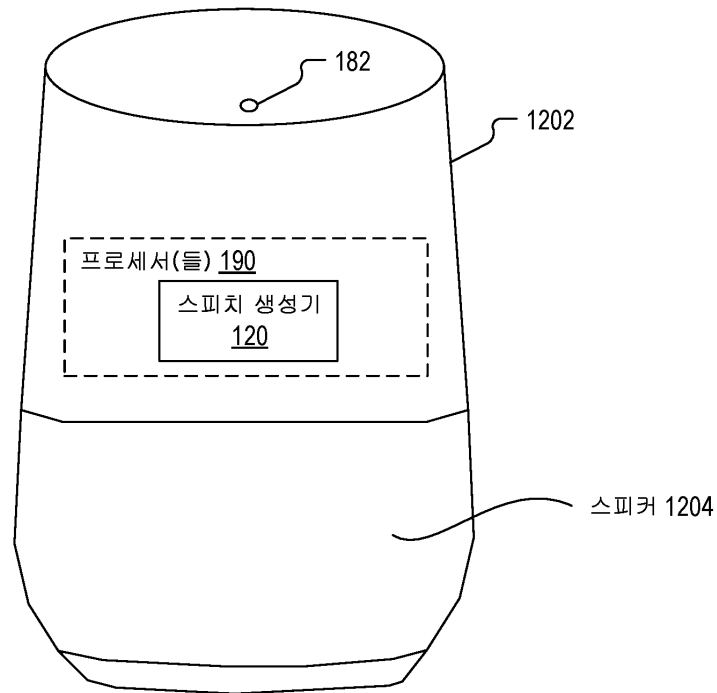
도면11

1100



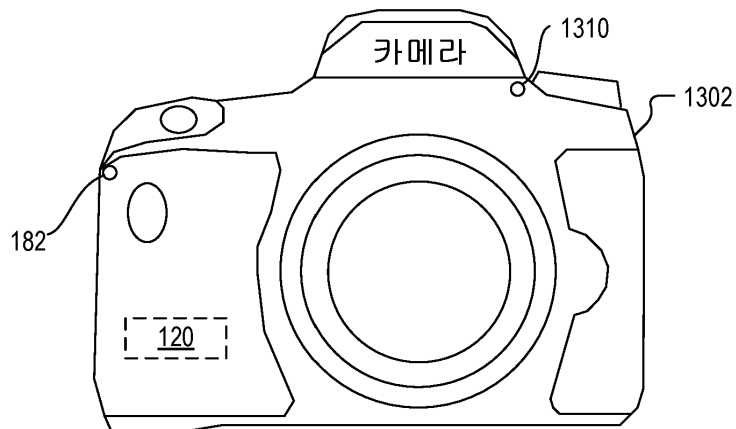
도면12

1200



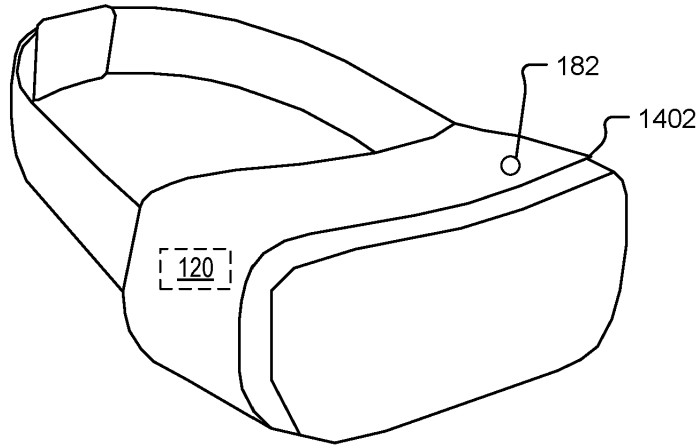
도면13

1300



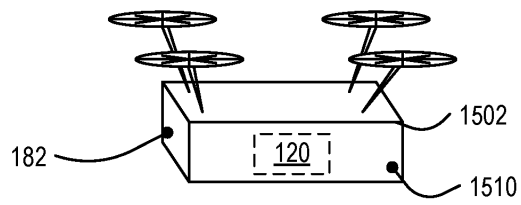
도면14

1400

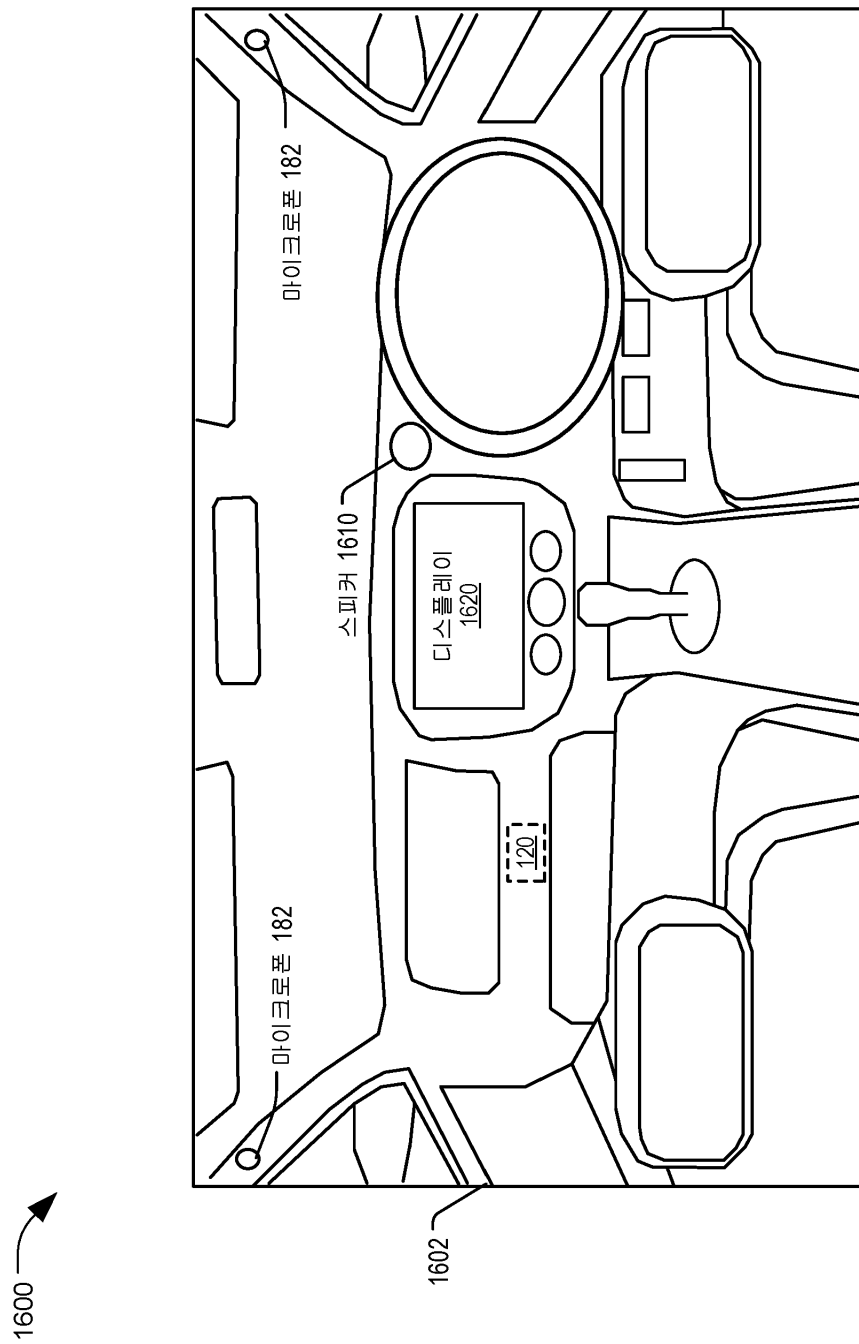


도면15

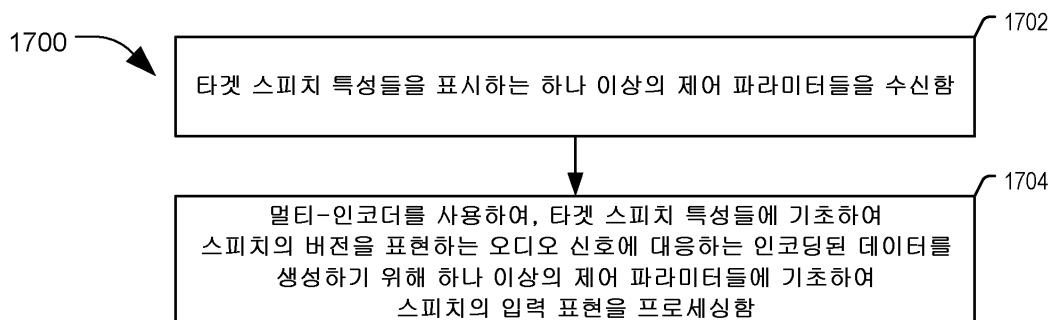
1500



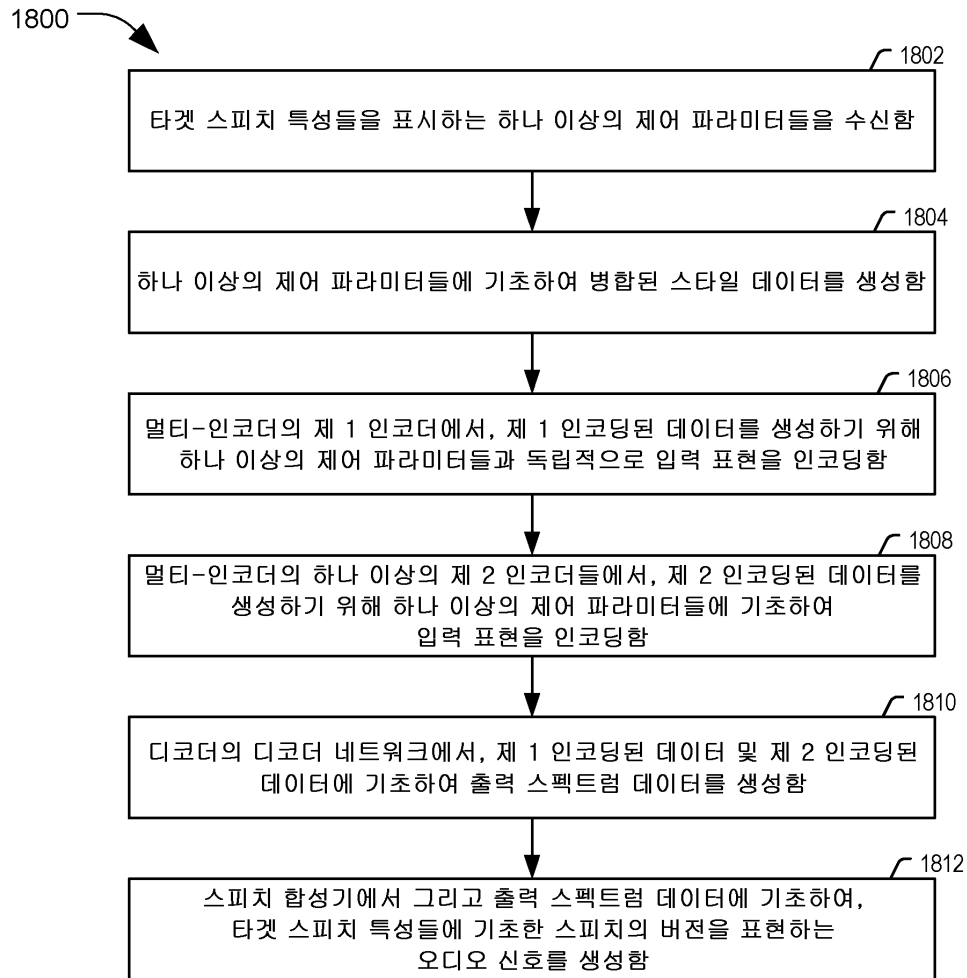
도면16



도면17



도면18



도면19

