(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0038504
(43) 공개일자 2024년03월25일

(51) 국제특허분류(Int. Cl.)
G10L 13/08 (2006.01) *G06N 20/00* (2019.01)
G10L 13/027 (2013.01) *G10L 13/06* (2006.01)
(52) CPC특허분류
G10L 13/08 (2013.01)
G06N 20/00 (2021.08)
(21) 출원번호 10-2022-0117469
(22) 출원일자 2022년09월16일
심사청구일자 2022년09월16일

(71) 출원인
네이버 주식회사
경기도 성남시 분당구 정자일로 95(정자동, 네이버 1784)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
송은우
경기도 성남시 분당구 정자일로 95, NAVER 1784
오수현
경기도 성남시 분당구 정자일로 95, NAVER 1784
(뒷면에 계속)
(74) 대리인
김한솔, 김세환, 김준식, 안제성

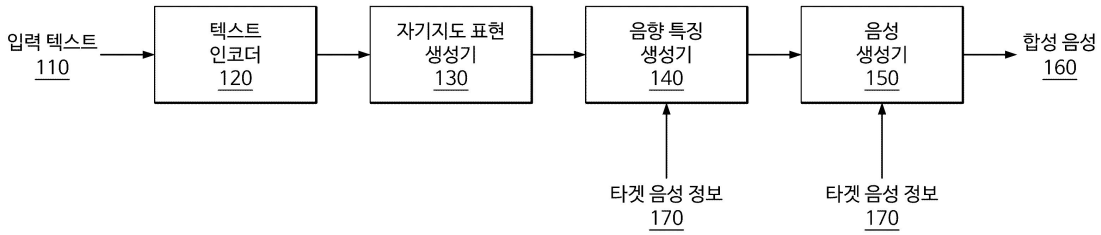
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **음성 합성 방법 및 시스템**

(57) 요 약

본 개시는 적어도 하나의 프로세서에 의해 수행되는 음성 합성 방법에 관한 것이다. 음성 합성 방법은, 입력 텍스트를 수신하는 단계, 텍스트 인코더를 이용하여, 입력 텍스트로부터 텍스트 표현을 생성하는 단계, 자기지도 표현 생성기를 이용하여, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하는 단계, 음향 특징 생성기를 이용하여, 자기지도 표현을 기초로 음향 특징을 생성하는 단계 및 음성 생성기를 이용하여, 음향 특징을 기초로 합성 음성을 생성하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G10L 13/027 (2013.01)
G10L 13/06 (2013.01)
(72) 발명자
이상훈
서울특별시 성북구 고려대로27길 34-8, b 305호
이성환
서울특별시 강남구 언주로30길 13, B동 2902호

명 세 서

청구범위

청구항 1

적어도 하나의 프로세서에 의해 수행되는, 음성 합성 방법에 있어서,

입력 텍스트를 수신하는 단계;

텍스트 인코더를 이용하여, 상기 입력 텍스트로부터 텍스트 표현을 생성하는 단계;

자기지도 표현 생성기를 이용하여, 상기 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하는 단계;

음향 특징 생성기를 이용하여, 상기 자기지도 표현을 기초로 음향 특징을 생성하는 단계; 및

음성 생성기를 이용하여, 상기 음향 특징을 기초로 합성 음성을 생성하는 단계

를 포함하는, 음성 합성 방법.

청구항 2

제1항에 있어서,

타겟 음성 정보를 수신하는 단계

를 더 포함하고,

상기 음향 특징을 생성하는 단계는,

상기 음향 특징 생성기를 이용하여, 상기 자기지도 표현 및 상기 타겟 음성 정보를 기초로 상기 음향 특징을 생성하고,

상기 합성 음성을 생성하는 단계는,

상기 음성 생성기를 이용하여, 상기 음향 특징 및 상기 타겟 음성 정보를 기초로 상기 합성 음성을 생성하는,

음성 합성 방법.

청구항 3

제1항에 있어서,

상기 음향 특징 생성기는, 학습된 음향 특징 제거기가 역변환된 모델이고,

상기 자기지도 표현 생성기는, 학습된 자기지도 표현 제거기가 역변환된 모델이고,

상기 자기지도 표현 제거기, 상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제1 학습 음성 또는 상기 제1 학습 음성에 대응하는 제1 학습 텍스트 중 적어도 하나에 기초하여 학습되는, 음성 합성 방법.

청구항 4

제3항에 있어서,

상기 자기지도 표현 제거기, 상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제1 손실에 기초하여 학습되고,

상기 제1 손실은 학습 합성 음성 및 제1 학습 음성에 기초하여 산출되고,

상기 학습 합성 음성은, 음향 특징 인코더 및 상기 음성 생성기를 이용하여 상기 제1 학습 음성을 기초로 생성되는, 음성 합성 방법.

청구항 5

제3항에 있어서,

상기 자기지도 표현 제거기, 상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제2 손실에 기초하여 학습되고,

상기 제2 손실은, 제1 학습 자기지도 표현 및 제2 학습 자기지도 표현에 기초하여 산출되고,

상기 제1 학습 자기지도 표현은, 음향 특징 인코더 및 상기 음향 특징 제거기를 이용하여, 상기 제1 학습 음성을 기초로 생성되고,

상기 제2 학습 자기지도 표현은, 자기지도 학습 모델 및 언어 특징 인코더를 이용하여 상기 제1 학습 음성으로부터 생성되는, 음성 합성 방법.

청구항 6

제5항에 있어서,

상기 자기지도 학습 모델은, 미리 학습된 음성 모델의 적어도 하나의 레이어를 포함하고,

상기 미리 학습된 음성 모델의 적어도 하나의 레이어는, 학습 과정에서 프리징(freezing)되는, 음성 합성 방법.

청구항 7

제3항에 있어서,

상기 자기지도 표현 제거기, 상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제3 손실에 기초하여 학습되고,

상기 제3 손실은, 제1 학습 텍스트 표현 및 제2 학습 텍스트 표현에 기초하여 산출되고,

상기 제1 학습 텍스트 표현은, 자기지도 학습 모델, 언어 특징 인코더 및 상기 자기지도 표현 제거기를 이용하여, 상기 제1 학습 음성으로부터 생성되고,

상기 제2 학습 텍스트 표현은, 상기 텍스트 인코더를 이용하여, 상기 제1 학습 텍스트로부터 생성되는, 음성 합성 방법.

청구항 8

제7항에 있어서,

상기 자기지도 학습 모델은, 미리 학습된 음성 모델의 적어도 하나의 레이어를 포함하고,

상기 미리 학습된 음성 모델의 적어도 하나의 레이어는, 학습 과정에서 프리징되는, 음성 합성 방법.

청구항 9

제3항에 있어서,

상기 자기지도 표현 제거기, 상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제4 손실에 기초하여 학습되고,

상기 제4 손실은, 예측된 단위 텍스트 및 상기 제1 학습 텍스트에 기초하여 산출되고,

상기 예측된 단위 텍스트는, 자기지도 학습 모델, 언어 특징 인코더 및 텍스트 예측기를 이용하여, 상기 제1 학습 음성으로부터 예측되는, 음성 합성 방법.

청구항 10

제9항에 있어서,

상기 예측된 단위 텍스트는 상기 제1 학습 텍스트에 포함되는 것으로 예측되는 음소(phoneme), 글자(character), 음절 또는 단어를 포함하는, 음성 합성 방법.

청구항 11

제3항에 있어서,

상기 음성 생성기 및 상기 음향 특징 제거기는 새로운 화자에 대한 제2 학습 음성을 기초로 전이 학습(fine-tuning) 되는, 음성 합성 방법.

청구항 12

제11항에 있어서,

상기 음향 특징을 생성하는 단계는,

상기 제2 학습 음성에 기초하여 전이 학습된 상기 음향 특징 제거기가 역변환된 음향 특징 생성기를 이용하여, 상기 자기지도 표현을 기초로 상기 음향 특징을 생성하고,

상기 합성 음성을 생성하는 단계는,

상기 제2 학습 음성에 기초하여 전이 학습된 음성 생성기를 이용하여, 상기 음향 특징을 기초로 상기 새로운 화자의 음성 특징을 반영하는 합성 음성을 생성하는, 음성 합성 방법.

청구항 13

적어도 하나의 프로세서에 의해 수행되는, 음성 합성 방법에 있어서,

소스 음성, 소스 음성 정보 및 타겟 음성 정보를 수신하는 단계;

음향 특징 인코더를 이용하여, 상기 소스 음성 및 상기 소스 음성 정보를 기초로 소스 음향 특징을 생성하는 단계;

음향 특징 제거기를 이용하여, 상기 소스 음향 특징 및 상기 소스 음성 정보를 기초로 언어적 정보를 포함하는 자기지도 표현을 생성하는 단계;

음향 특징 생성기를 이용하여, 상기 자기지도 표현 및 상기 타겟 음성 정보를 기초로 타겟 음향 특징을 생성하는 단계; 및

음성 생성기를 이용하여, 상기 타겟 음향 특징 및 상기 타겟 음성 정보를 기초로 변환 음성을 생성하는 단계

를 포함하는, 음성 합성 방법.

청구항 14

제13항에 있어서,

상기 음향 특징 생성기는, 학습된 음향 특징 제거기가 역변환된 모델이고,

상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제3 학습 음성에 기초하여 학습되는, 음성 합성 방법.

청구항 15

제14항에 있어서,

상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제5 손실에 기초하여 학습되고,

상기 제5 손실은 학습 합성 음성 및 제3 학습 음성에 기초하여 산출되고,

상기 학습 합성 음성은, 상기 음향 특징 인코더 및 상기 음성 생성기를 이용하여, 상기 제3 학습 음성을 기초로 생성되는, 음성 합성 방법.

청구항 16

제14항에 있어서,

상기 음향 특징 제거기 또는 상기 음성 생성기 중 적어도 하나는, 제6 손실에 기초하여 학습되고,

상기 제6 손실은, 제1 학습 자기지도 표현 및 제2 학습 자기지도 표현에 기초하여 산출되고,

상기 제1 학습 자기지도 표현은, 상기 음향 특징 인코더 및 상기 음향 특징 제거기를 이용하여, 상기 제3 학습 음성을 기초로 생성되고,

상기 제2 학습 자기지도 표현은, 자기지도 학습 모델 및 언어 특징 인코더를 이용하여 상기 제3 학습 음성으로부터 생성되는, 음성 합성 방법.

청구항 17

제16항에 있어서,

상기 자기지도 학습 모델은, 미리 학습된 음성 모델의 적어도 하나의 레이어를 포함하고,

상기 미리 학습된 음성 모델의 적어도 하나의 레이어는, 학습 과정에서 프리징되는, 음성 합성 방법.

청구항 18

제1항 내지 제17항 중 어느 한 항에 따른 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능 기록 매체에 저장된 컴퓨터 프로그램.

청구항 19

정보 처리 시스템으로서,

메모리; 및

상기 메모리에 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로그램은,

입력 텍스트를 수신하고,

텍스트 인코더를 이용하여, 상기 입력 텍스트로부터 텍스트 표현을 생성하고,

자기지도 표현 생성기를 이용하여, 상기 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하고,

음향 특징 생성기를 이용하여, 상기 자기지도 표현을 기초로 음향 특징을 생성하고,

음성 생성기를 이용하여, 상기 음향 특징을 기초로 합성 음성을 생성하기 위한 명령어들을 포함하는, 정보 처리 시스템.

청구항 20

정보 처리 시스템으로서,

메모리; 및

상기 메모리에 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로그램은,

소스 음성, 소스 음성 정보 및 타겟 음성 정보를 수신하고,

음향 특징 인코더를 이용하여, 상기 소스 음성 및 상기 소스 음성 정보를 기초로 소스 음향 특징을 생성하고,

음향 특징 제거기를 이용하여, 상기 소스 음향 특징 및 상기 소스 음성 정보를 기초로 자기지도 표현을 생성하고,

음향 특징 생성기를 이용하여, 상기 자기지도 표현 및 상기 타겟 음성 정보를 기초로 타겟 음향 특징을 생성하고,

음성 생성기를 이용하여, 상기 타겟 음향 특징 및 상기 타겟 음성 정보를 기초로 변환 음성을 생성하기 위한 명령어들을 포함하는, 정보 처리 시스템.

발명의 설명

기 술 분 야

[0001] 본 개시는 음성 합성 방법 및 시스템에 관한 것으로, 구체적으로, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하고, 자기지도 표현에 기초하여 음향 특징을 생성함으로써, 언어적 정보가 유실되지 않도록 합성 음성을 생성하는 방법 및 시스템에 관한 것이다.

배 경 기 술

[0003] 음성 합성 기술은 안내방송, 내레이션, 네비게이션, 인공지능 비서 등과 같이, 사람의 음성이 필요한 애플리케이션에서 실제 사람의 음성을 사전에 녹음해 두지 않고 필요한 음성을 재생하기 위해 사용되는 기술이다. 최근에는 인공신경망(artificial neural network) 기반의 음성 합성 방법이 활발히 연구되고 있으며, 이 방법에 따라 합성된 음성은 기존의 방법에 비해 훨씬 자연스러운 음성 특징을 반영하고 있다.

[0004] 대부분의 인공신경망 기반 음성 합성 방법은 텍스트 표현으로부터 음향 특징을 생성하고, 이를 기초로 합성 음성을 생성할 수 있다. 다만, 이러한 종래의 방법에 따르면, 텍스트 표현으로부터 곧바로 음향 특징을 생성함에 따라 언어적 정보가 유실되어, 발음이 뭉개지거나 지나치게 매끄러운 합성 음성이 생성될 수 있다. 즉, 종래의 방법에 따라 생성된 합성 음성은 텍스트의 언어적 특징을 충분히 반영하지 못할 수 있다.

발명의 내용

해결하려는 과제

[0006] 본 개시는 상기와 같은 문제점을 해결하기 위한 음성 합성 방법, 기록매체에 저장된 컴퓨터 프로그램 및 장치(시스템)를 제공한다.

과제의 해결 수단

[0008] 본 개시는 방법, 장치(시스템) 또는 판독 가능 저장 매체에 저장된 컴퓨터 프로그램을 포함한 다양한 방식으로 구현될 수 있다.

[0009] 본 개시의 일 실시예에 따르면, 적어도 하나의 프로세서에 의해 수행되는, 음성 합성 방법은, 입력 텍스트를 수신하는 단계, 텍스트 인코더를 이용하여, 입력 텍스트로부터 텍스트 표현을 생성하는 단계, 자기지도 표현 생성기를 이용하여, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하는 단계, 음향 특징 생성기를 이용하여, 자기지도 표현을 기초로 음향 특징을 생성하는 단계 및 음성 생성기를 이용하여, 음향 특징을 기초로 합성 음성을 생성하는 단계를 포함한다.

[0010] 본 개시의 일 실시예에 따르면, 적어도 하나의 프로세서에 의해 수행되는, 음성 합성 방법은, 소스 음성, 소스 음성 정보 및 타겟 음성 정보를 수신하는 단계, 음향 특징 인코더를 이용하여, 소스 음성 및 소스 음성 정보를 기초로 소스 음향 특징을 생성하는 단계, 음향 특징 제거기를 이용하여, 소스 음향 특징 및 소스 음성 정보를 기초로 언어적 정보를 포함하는 자기지도 표현을 생성하는 단계, 음향 특징 생성기를 이용하여, 자기지도 표현 및 타겟 음성 정보를 기초로 타겟 음향 특징을 생성하는 단계 및 음성 생성기를 이용하여, 타겟 음향 특징 및 타겟 음성 정보를 기초로 변환 음성을 생성하는 단계를 포함한다.

[0011] 본 개시의 일 실시예에 따른 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능 기록 매체에 저장된 컴퓨터 프로그램이 제공된다.

[0012] 본 개시의 일 실시예에 따른 정보 처리 시스템은, 메모리 및 메모리에 연결되고, 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 적어도 하나의 프로그램은, 입력 텍스트를 수신하고, 텍스트 인코더를 이용하여, 입력 텍스트로부터 텍스트 표현을 생성하고, 자기지도 표현 생성기를 이용하여, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하고, 음향 특징 생성기를 이용하여, 자기지도 표현을 기초로 음향 특징을 생성하고, 음성 생성기를 이용하여, 음향 특징을 기초로 합성 음성을 생성하기 위한 명령어들을 포함한다.

[0013] 본 개시의 일 실시예에 따른 정보 처리 시스템은, 메모리 및 메모리에 연결되고, 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 적어도 하나의 프로그램은, 소스 음성, 소스 음성 정보 및 타겟 음성 정보를 수신하고, 음향 특징 인코더를 이용하여, 소스 음성 및 소스 음성 정보를 기초로 소스 음향 특징을 생성하고, 음향 특징 제거기를 이용하여, 소스 음향 특징 및 소스 음성 정보를 기초로 자기지도 표현을 생성하고, 음향 특징 생성기를 이용하여, 자기지도 표현 및 타겟 음성 정보를 기초로 타겟 음향 특징을 생성하고, 음성 생성기를 이용하여, 타겟 음향 특징 및 타겟 음성 정보를 기초로 변환 음성을 생성하기 위한 명령어들을 포함한다.

발명의 효과

[0015] 본 개시의 일부 실시예에 따르면, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하고 이에 기초하여 음향 특징을 생성하는 2단계의 구조를 채택하고, 언어적 특성 및 음향적 특성을 계층적으로 학습함으로써, 언어적 특징을 충분히 반영하는 정확한 발음의 합성 음성이 생성될 수 있다.

[0016] 본 개시의 일부 실시예에 따르면, 음성의 언어적 정보를 유지하면서 음향 특징만을 제거하도록 학습된 음향 특징 제거기 및 학습된 음향 특징 제거기가 역변환된 음향 특징 생성기를 이용하여, 소스 음성의 언어적 정보는 그대로 유지하면서 음향 특징만 변환된 변환 음성이 생성될 수 있다.

[0017] 본 개시의 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급되지 않은 다른 효과들은 청구범위의 기재로부

터 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자(“통상의 기술자” 라 함)에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0019] 본 개시의 실시예들은, 이하 설명하는 첨부 도면들을 참조하여 설명될 것이며, 여기서 유사한 참조 번호는 유사한 요소들을 나타내지만, 이에 한정되지는 않는다.

도 1은 본 개시의 일 실시예에 따른 합성 음성을 생성하는 방법의 예시를 나타내는 도면이다.

도 2는 본 개시의 일 실시예에 따른 정보 처리 시스템이 복수의 사용자 단말과 통신 가능하도록 연결된 구성을 나타내는 개요도이다.

도 3은 본 개시의 일 실시예에 따른 사용자 단말 및 정보 처리 시스템의 내부 구성을 나타내는 블록도이다.

도 4는 본 개시의 일 실시예에 따른 음성 합성 모델을 학습시키는 방법의 예시를 나타내는 도면이다.

도 5는 본 개시의 일 실시예에 따른 학습된 음성 합성 모델을 이용하여 합성 음성을 생성하는 방법의 예시를 나타내는 도면이다.

도 6은 본 개시의 일 실시예에 따른 변환 음성을 생성하는 방법의 예시를 나타내는 도면이다.

도 7은 본 개시의 일 실시예에 따라 새로운 화자에 대한 학습 음성을 이용하여 음성 합성 모델을 전이 학습시키는 방법의 예시를 나타내는 도면이다.

도 8은 본 개시의 일 실시예에 따른 음성 합성 모델의 성능을 평가한 예시를 나타내는 도면이다.

도 9는 본 개시의 일 실시예에 따른 음성 합성 방법의 예시를 나타내는 흐름도이다.

도 10은 본 개시의 일 실시예에 따른 음성 합성 방법의 예시를 나타내는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 이하, 본 개시의 실시를 위한 구체적인 내용을 첨부된 도면을 참조하여 상세히 설명한다. 다만, 이하의 설명에서는 본 개시의 요지를 불필요하게 흐릴 우려가 있는 경우, 널리 알려진 기능이나 구성에 관한 구체적 설명은 생략하기로 한다.

[0021] 첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응되는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나, 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지는 않는다.

[0022] 개시된 실시예의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시가 완전하도록 하고, 본 개시가 통상의 기술자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이다.

[0023] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 개시된 실시예에 대해 구체적으로 설명하기로 한다. 본 명세서에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 관련 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0024] 본 명세서에서의 단수의 표현은 문맥상 명백하게 단수인 것으로 특정하지 않는 한, 복수의 표현을 포함한다. 또한, 복수의 표현은 문맥상 명백하게 복수인 것으로 특정하지 않는 한, 단수의 표현을 포함한다. 명세서 전체에서 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다.

[0025] 또한, 명세서에서 사용되는 '모듈' 또는 '부'라는 용어는 소프트웨어 또는 하드웨어 구성요소를 의미하며, '모듈' 또는 '부'는 어떤 역할들을 수행한다. 그렇지만, '모듈' 또는 '부'는 소프트웨어 또는 하드웨어에 한정되

는 의미는 아니다. '모듈' 또는 '부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서, '모듈' 또는 '부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 또는 변수들 중 적어도 하나를 포함할 수 있다. 구성요소들과 '모듈' 또는 '부'들은 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '모듈' 또는 '부'들로 결합되거나 추가적인 구성요소들과 '모듈' 또는 '부'들로 더 분리될 수 있다.

[0026] 본 개시의 일 실시예에 따르면, '모듈' 또는 '부'는 프로세서 및 메모리로 구현될 수 있다. '프로세서'는 범용 프로세서, 중앙 처리 장치(CPU), 마이크로프로세서, 디지털 신호 프로세서(DSP), 제어기, 마이크로제어기, 상태 머신 등을 포함하도록 넓게 해석되어야 한다. 몇몇 환경에서, '프로세서'는 주문형 반도체(ASIC), 프로그램가능 로직 디바이스(PLD), 필드 프로그램가능 게이트 어레이(FPGA) 등을 지칭할 수도 있다. '프로세서'는, 예를 들어, DSP와 마이크로프로세서의 조합, 복수의 마이크로프로세서들의 조합, DSP 코어와 결합한 하나 이상의 마이크로프로세서들의 조합, 또는 임의의 다른 그러한 구성들의 조합과 같은 처리 디바이스들의 조합을 지칭할 수도 있다. 또한, '메모리'는 전자 정보를 저장 가능한 임의의 전자 컴포넌트를 포함하도록 넓게 해석되어야 한다. '메모리'는 임의 액세스 메모리(RAM), 판독-전용 메모리(ROM), 비-휘발성 임의 액세스 메모리(NVRAM), 프로그램가능 판독-전용 메모리(PROM), 소거-프로그램가능 판독 전용 메모리(EPROM), 전기적으로 소거가능 PROM(EEPROM), 플래쉬 메모리, 자기 또는 광학 데이터 저장장치, 레지스터들 등과 같은 프로세서-판독가능 매체의 다양한 유형들을 지칭할 수도 있다. 프로세서가 메모리로부터 정보를 판독하고/하거나 메모리에 정보를 기록할 수 있다면 메모리는 프로세서와 전자 통신 상태에 있다고 불린다. 프로세서에 집적된 메모리는 프로세서와 전자 통신 상태에 있다.

[0027] 본 개시에서, '시스템'은 서버 장치와 클라우드 장치 중 적어도 하나의 장치를 포함할 수 있으나, 이에 한정되는 것은 아니다. 예를 들어, 시스템은 하나 이상의 서버 장치로 구성될 수 있다. 다른 예로서, 시스템은 하나 이상의 클라우드 장치로 구성될 수 있다. 또 다른 예로서, 시스템은 서버 장치와 클라우드 장치가 함께 구성되어 동작될 수 있다.

[0028] 본 개시에서, '기계학습 모델'은 주어진 입력에 대한 해답(answer)을 추론하는데 사용하는 임의의 모델을 포함할 수 있다. 일 실시예에 따르면, 기계학습 모델은 입력 레이어(층), 복수 개의 은닉 레이어 및 출력 레이어를 포함한 인공신경망 모델을 포함할 수 있다. 여기서, 각 레이어는 복수의 노드를 포함할 수 있다. 본 개시에서, 자기지도 표현 생성기, 음향 특징 생성기, 음성 생성기, 텍스트 예측기 등 복수의 기계학습 모델을 별도의 기계학습 모델로서 설명하나, 이에 한정되지 않으며, 복수의 기계학습 모델의 일부 또는 전체는 하나의 기계학습 모델로 구현될 수 있다. 또한, 단일 기계학습 모델로서 설명되는 모델은 복수의 기계학습 모델로서 구현될 수도 있다. 본 개시에서, 기계학습 모델은 인공신경망 모델을 지칭할 수 있으며, 인공신경망 모델은 기계학습 모델을 지칭할 수 있다.

[0029] 본 개시에서, '복수의 A의 각각' 또는 '복수의 A 각각'은 복수의 A에 포함된 모든 구성 요소의 각각을 지칭하거나, 복수의 A에 포함된 일부 구성 요소의 각각을 지칭할 수 있다.

[0030] 도 1은 본 개시의 일 실시예에 따른 합성 음성(160)을 생성하는 방법의 예시를 나타내는 도면이다. 일 실시예에 따르면, 정보 처리 시스템은 입력 텍스트(110)를 수신하고, 음성 합성 모델을 이용하여 입력 텍스트(110)를 기초로 합성 음성(160)을 생성할 수 있다. 예를 들어, 입력 텍스트(110)는 텍스트에 포함된 단위 텍스트들의 집합일 수 있다. 구체적 예로, 입력 텍스트(110)는 복수의 문자를 포함하는 문자열 또는 복수의 음소를 포함하는 음소열 등일 수 있다. 또한, 합성 음성(160)은 텍스트가 화자에 의해 발화되는 것과 같이 들리는 음성일 수 있다.

[0031] 일 실시예에 따르면, 음성 합성 모델은 텍스트 인코더(120), 자기지도 표현 생성기(130), 음향 특징 생성기(140) 및 음성 생성기(150)를 포함할 수 있다.

[0032] 텍스트 인코더(120)는 입력 텍스트(110)를 입력으로 수신하고, 입력 텍스트(110)로부터 텍스트 정보를 포함하는 텍스트 표현(text representation)을 생성할 수 있다. 예를 들어, 텍스트 표현은 입력 텍스트(110)의 텍스트 정보를 포함하는 임베딩 벡터일 수 있다.

[0033] 자기지도 표현 생성기(130)는 텍스트 인코더(120)에 의해 생성된 텍스트 표현을 입력으로 수신하고, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현(self-supervised representation)을 생성할 수 있다.

[0034] 음향 특징 생성기(140)는 자기지도 표현 생성기(130)에 의해 생성된 자기지도 표현을 입력으로 수신하고, 자기지도 표현을 기초로 음향 특징을 생성할 수 있다. 일 실시예에서, 음성 합성 모델이 다화자(multi-speaker) 모델인 경우, 음향 특징 생성기(140)는 타겟 음성 정보(170)를 추가적으로 수신하고, 자기지도 표현 및 타겟 음성 정보(170)를 기초로 음향 특징을 생성할 수 있다.

[0035] 음성 생성기(150)는 음향 특징 생성기(140)에 의해 생성된 음향 특징을 입력으로 수신하고, 음향 특징을 기초로 합성 음성(160)을 생성할 수 있다. 일 실시예에서, 음성 합성 모델이 다화자 모델인 경우, 음성 생성기(150)는 타겟 음성 정보(170)를 추가적으로 수신하고, 음향 특징 및 타겟 음성 정보(170)를 기초로 타겟 음성 정보를 반영하는 합성 음성(160)을 생성할 수 있다.

[0036] 종래의 종단간 음성 합성 방법에 따르면, 텍스트 표현으로부터 음향 특징을 생성하고, 이를 기초로 합성 음성을 생성하였다. 이러한 종래의 방법에 따르면, 텍스트 표현으로부터 곧바로 음향 특징을 생성함에 따라 언어적 정보가 유실되어, 발음이 뭉개지거나 지나치게 매끄러운 합성 음성이 생성되는 등 생성된 합성 음성이 텍스트의 언어적 특징을 충분히 반영하지 못한다는 문제점이 있었다. 본 개시의 일 실시예에 따르면, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성하고 이에 기초하여 음향 특징을 생성하는 2단계의 구조를 채택하고, 언어적 특성 및 음향적 특성을 계층적으로 학습함으로써, 언어적 특징을 충분히 반영하는 정확한 발음의 합성 음성(160)을 생성할 수 있다.

[0037] 본 개시에 따른 음성 합성 방법은 정보 처리 시스템, 사용자 단말 등 임의의 수행 주체에 의해 수행될 수 있다. 다만, 설명의 편의를 위해 이하에서는 정보 처리 시스템(예를 들어, 정보 처리 시스템의 적어도 하나의 프로세서)에 의해 음성 합성 방법에 수행되는 것으로 가정하고 서술하고자 한다.

[0038] 도 2는 본 개시의 일 실시예에 따른 정보 처리 시스템(230)이 복수의 사용자 단말(210_1, 210_2, 210_3)과 통신 가능하도록 연결된 구성을 나타내는 개요도이다. 도시된 바와 같이, 복수의 사용자 단말(210_1, 210_2, 210_3)은 네트워크(220)를 통해 음성 합성 서비스 및/또는 음성 변환 서비스를 제공할 수 있는 정보 처리 시스템(230)과 연결될 수 있다. 여기서, 복수의 사용자 단말(210_1, 210_2, 210_3)은 음성 합성 서비스 및/또는 음성 변환 서비스를 제공받을 사용자의 단말을 포함할 수 있다. 일 실시예에서, 정보 처리 시스템(230)은 음성 합성 서비스 및/또는 음성 변환 서비스 등과 관련된 컴퓨터 실행 가능한 프로그램(예를 들어, 다운로드 가능한 애플리케이션) 및 데이터를 저장, 제공 및 실행할 수 있는 하나 이상의 서버 장치 및/또는 데이터베이스, 또는 클라우드 컴퓨팅 서비스 기반의 하나 이상의 분산 컴퓨팅 장치 및/또는 분산 데이터베이스를 포함할 수 있다.

[0039] 정보 처리 시스템(230)에 의해 제공되는 음성 합성 서비스 및/또는 음성 변환 서비스는, 복수의 사용자 단말(210_1, 210_2, 210_3)의 각각에 설치된 음성 합성 애플리케이션, 음성 변환 애플리케이션, 모바일 브라우저 애플리케이션 또는 웹 브라우저 등을 통해 사용자에게 제공될 수 있다. 예를 들어, 정보 처리 시스템(230)은 음성 합성 애플리케이션 등을 통해 사용자 단말(210_1, 210_2, 210_3)로부터 수신되는 텍스트에 대한 합성 음성 요청에 대응하는 정보를 제공하거나 대응하는 처리를 수행할 수 있다.

[0040] 복수의 사용자 단말(210_1, 210_2, 210_3)은 네트워크(220)를 통해 정보 처리 시스템(230)과 통신할 수 있다. 네트워크(220)는 복수의 사용자 단말(210_1, 210_2, 210_3)과 정보 처리 시스템(230) 사이의 통신이 가능하도록 구성될 수 있다. 네트워크(220)는 설치 환경에 따라, 예를 들어, 이더넷(Ethernet), 유선 홈 네트워크(Power Line Communication), 전화선 통신 장치 및 RS-serial 통신 등의 유선 네트워크, 이동통신망, WLAN(Wireless LAN), Wi-Fi, Bluetooth 및 ZigBee 등과 같은 무선 네트워크 또는 그 조합으로 구성될 수 있다. 통신 방식은 제한되지 않으며, 네트워크(220)가 포함할 수 있는 통신망(일례로, 이동통신망, 유선 인터넷, 무선 인터넷, 방송망, 위성망 등)을 활용하는 통신 방식뿐만 아니라 사용자 단말(210_1, 210_2, 210_3) 사이의 근거리 무선 통신 역시 포함될 수 있다.

[0041] 도 2에서 휴대폰 단말(210_1), 태블릿 단말(210_2) 및 PC 단말 (210_3)이 사용자 단말의 예로서 도시되었으나, 이에 한정되지 않으며, 사용자 단말(210_1, 210_2, 210_3)은 유선 및/또는 무선 통신이 가능하고 음성 합성 애플리케이션, 음성 변환 애플리케이션, 모바일 브라우저 애플리케이션 또는 웹 브라우저 등이 설치되어 실행될 수 있는 임의의 컴퓨팅 장치일 수 있다. 예를 들어, 사용자 단말은, AI 스피커, 스마트폰, 휴대폰, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스, VR(virtual reality) 디바이스, AR(augmented reality) 디바이스, 셋톱 박스 등을 포함할 수 있다. 또한, 도 2에는 3개의 사용자 단말(210_1, 210_2, 210_3)이 네트워크(220)를 통해 정보 처리 시스템(230)과 통신하는 것으로 도시되어 있으나, 이에 한정되지 않으며, 상이한 수의 사용자 단말이

네트워크(220)를 통해 정보 처리 시스템(230)과 통신하도록 구성될 수도 있다.

[0042] 일 실시예에 따르면, 정보 처리 시스템(230)은 복수의 사용자 단말(210_1, 210_2, 210_3)로부터 입력 텍스트와 연관된 데이터를 수신할 수 있다. 그리고 나서, 정보 처리 시스템(230)은 이렇게 수신된 입력 텍스트와 연관된 데이터에 기초하여 합성 음성을 생성하여, 복수의 사용자 단말(210_1, 210_2, 210_3)에 제공할 수 있다. 추가적으로 또는 대안적으로, 정보 처리 시스템(230)은 복수의 사용자 단말(210_1, 210_2, 210_3)로부터 소스 음성과 연관된 데이터를 수신할 수 있다. 그리고 나서, 정보 처리 시스템(230)은 이렇게 수신된 소스 음성과 연관된 데이터에 기초하여 변환 음성을 생성하여, 복수의 사용자 단말(210_1, 210_2, 210_3)에 제공할 수 있다.

[0043] 도 3은 본 개시의 일 실시예에 따른 사용자 단말(210) 및 정보 처리 시스템(230)의 내부 구성을 나타내는 블록도이다. 사용자 단말(210)은 음성 합성 애플리케이션, 음성 변환 애플리케이션, 모바일 브라우저 애플리케이션 또는 웹 브라우저 등을 실행 가능하고 유/무선 통신이 가능한 임의의 컴퓨팅 장치를 지칭할 수 있으며, 예를 들어, 도 2의 휴대폰 단말(210_1), 태블릿 단말(210_2), PC 단말(210_3) 등을 포함할 수 있다. 도시된 바와 같이, 사용자 단말(210)은 메모리(312), 프로세서(314), 통신 모듈(316) 및 입출력 인터페이스(318)를 포함할 수 있다. 이와 유사하게, 정보 처리 시스템(230)은 메모리(332), 프로세서(334), 통신 모듈(336) 및 입출력 인터페이스(338)를 포함할 수 있다. 도 3에 도시된 바와 같이, 사용자 단말(210) 및 정보 처리 시스템(230)은 각각의 통신 모듈(316, 336)을 이용하여 네트워크(220)를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다. 또한, 입출력 장치(320)는 입출력 인터페이스(318)를 통해 사용자 단말(210)에 정보 및/또는 데이터를 입력하거나 사용자 단말(210)로부터 생성된 정보 및/또는 데이터를 출력하도록 구성될 수 있다.

[0044] 메모리(312, 332)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(312, 332)는 ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 다른 예로서, ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리와는 구분되는 별도의 영구 저장 장치로서 사용자 단말(210) 또는 정보 처리 시스템(230)에 포함될 수 있다. 또한, 메모리(312, 332)에는 운영체제와 적어도 하나의 프로그램 코드(예를 들어, 사용자 단말(210)에 설치되어 구동되는 음성 합성 애플리케이션 등을 위한 코드)가 저장될 수 있다.

[0045] 이러한 소프트웨어 구성요소들은 메모리(312, 332)와는 별도의 컴퓨터에서 판독가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독가능한 기록매체는 이러한 사용자 단말(210) 및 정보 처리 시스템(230)에 직접 연결가능한 기록 매체를 포함할 수 있는데, 예를 들어, 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 예로서, 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 모듈을 통해 메모리(312, 332)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 애플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 네트워크(220)를 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(312, 332)에 로딩될 수 있다.

[0046] 프로세서(314, 334)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(312, 332) 또는 통신 모듈(316, 336)에 의해 프로세서(314, 334)로 제공될 수 있다. 예를 들어, 프로세서(314, 334)는 메모리(312, 332)와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.

[0047] 통신 모듈(316, 336)은 네트워크(220)를 통해 사용자 단말(210)과 정보 처리 시스템(230)이 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 사용자 단말(210) 및/또는 정보 처리 시스템(230)이 다른 사용자 단말 또는 다른 시스템(일례로 별도의 클라우드 시스템 등)과 통신하기 위한 구성 또는 기능을 제공할 수 있다. 일례로, 사용자 단말(210)의 프로세서(314)가 메모리(312) 등과 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청 또는 데이터(예를 들어, 입력 텍스트에 대한 합성 음성 요청, 소스 음성에 대한 변환 음성 요청 등)는 통신 모듈(316)의 제어에 따라 네트워크(220)를 통해 정보 처리 시스템(230)으로 전달될 수 있다. 역으로, 정보 처리 시스템(230)의 프로세서(334)의 제어에 따라 제공되는 제어 신호나 명령이 통신 모듈(336)과 네트워크(220)를 거쳐 사용자 단말(210)의 통신 모듈(316)을 통해 사용자 단말(210)에 수신될 수 있다. 예를 들어, 사용자 단말(210)은 정보 처리 시스템(230)으로부터 통신 모듈(316)을 통해 입력 텍스트에 대한 합성 음성 또는 소스 음성에 대한 변환 음성 등을 수신할 수 있다.

[0048] 입출력 인터페이스(318)는 입출력 장치(320)와의 인터페이스를 위한 수단일 수 있다. 일 예로서, 입력 장치는 오디오 센서 및/또는 이미지 센서를 포함한 카메라, 키보드, 마이크로폰, 마우스 등의 장치를, 그리고 출력 장

치는 디스플레이, 스피커, 햅틱 피드백 디바이스(haptic feedback device) 등과 같은 장치를 포함할 수 있다. 다른 예로, 입출력 인터페이스(318)는 터치스크린 등과 같이 입력과 출력을 수행하기 위한 구성 또는 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 사용자 단말(210)의 프로세서(314)가 메모리(312)에 로딩된 컴퓨터 프로그램의 명령을 처리함에 있어서 정보 처리 시스템(230)이나 다른 사용자 단말이 제공하는 정보 및/또는 데이터를 이용하여 구성되는 서비스 화면 등이 입출력 인터페이스(318)를 통해 디스플레이에 표시될 수 있다. 도 3에서는 입출력 장치(320)가 사용자 단말(210)에 포함되지 않도록 도시되어 있으나, 이에 한정되지 않으며, 사용자 단말(210)과 하나의 장치로 구성될 수 있다. 또한, 정보 처리 시스템(230)의 입출력 인터페이스(338)는 정보 처리 시스템(230)과 연결되거나 정보 처리 시스템(230)이 포함할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 도 3에서는 입출력 인터페이스(318, 338)가 프로세서(314, 334)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(318, 338)가 프로세서(314, 334)에 포함되도록 구성될 수 있다.

[0049] 사용자 단말(210) 및 정보 처리 시스템(230)은 도 3의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 일 실시예에 따르면, 사용자 단말(210)은 상술된 입출력 장치(320) 중 적어도 일부를 포함하도록 구현될 수 있다. 또한, 사용자 단말(210)은 트랜시버(transceiver), GPS(Global Positioning system) 모듈, 카메라, 각종 센서, 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수 있다. 예를 들어, 사용자 단말(210)이 스마트폰인 경우, 일반적으로 스마트폰이 포함하고 있는 구성요소를 포함할 수 있으며, 예를 들어, 가속도 센서, 자이로 센서, 카메라 모듈, 각종 물리적인 버튼, 터치패널을 이용한 버튼, 입출력 포트, 진동을 위한 진동기 등의 다양한 구성요소들이 사용자 단말(210)에 더 포함되도록 구현될 수 있다. 일 실시예에 따르면, 사용자 단말(210)의 프로세서(314)는 음성 합성 서비스를 제공하는 애플리케이션 등이 동작하도록 구성될 수 있다. 이 때, 해당 애플리케이션 및/또는 프로그램과 연관된 코드가 사용자 단말(210)의 메모리(312)에 로딩될 수 있다.

[0050] 음성 합성 서비스를 제공하는 애플리케이션 등을 위한 프로그램이 동작되는 동안에, 프로세서(314)는 입출력 인터페이스(318)와 연결된 터치 스크린, 키보드, 오디오 센서 및/또는 이미지 센서를 포함한 카메라, 마이크로폰 등의 입력 장치를 통해 입력되거나 선택된 텍스트, 이미지, 영상, 음성 및/또는 동작 등을 수신할 수 있으며, 수신된 텍스트, 이미지, 영상, 음성 및/또는 동작 등을 메모리(312)에 저장하거나 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공할 수 있다. 예를 들어, 프로세서(314)는 입력 텍스트를 수신하여, 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공할 수 있다. 다른 예로서, 프로세서(314)는 소스 음성을 수신하여, 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공할 수 있다.

[0051] 사용자 단말(210)의 프로세서(314)는 입력 장치(320), 다른 사용자 단말, 정보 처리 시스템(230) 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다. 프로세서(314)에 의해 처리된 정보 및/또는 데이터는 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공될 수 있다. 사용자 단말(210)의 프로세서(314)는 입출력 인터페이스(318)를 통해 입출력 장치(320)로 정보 및/또는 데이터를 전송하여, 출력할 수 있다. 예를 들면, 프로세서(314)는 수신한 정보 및/또는 데이터를 사용자 단말의 화면에 디스플레이할 수 있다.

[0052] 정보 처리 시스템(230)의 프로세서(334)는 복수의 사용자 단말(210) 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다. 프로세서(334)에 의해 처리된 정보 및/또는 데이터는 통신 모듈(336) 및 네트워크(220)를 통해 사용자 단말(210)에 제공할 수 있다. 일 실시예에서, 정보 처리 시스템(230)의 프로세서(334)는 복수의 사용자 단말(210)로부터 수신된 입력 텍스트와 연관된 데이터에 기초하여, 언어적 정보를 표현하는 자기지도 표현 및 음향 특징을 생성하고, 이에 기초하여 합성 음성을 생성할 수 있다.

[0053] 정보 처리 시스템(230)의 프로세서(334)는 사용자 단말(210)의 디스플레이 출력 가능 장치(예: 터치 스크린, 디스플레이 등), 음성 출력 가능 장치(예: 스피커) 등의 출력 장치(320)를 통해 처리된 정보 및/또는 데이터를 출력하도록 구성될 수 있다. 예를 들어, 정보 처리 시스템(230)의 프로세서(334)는, 생성된 합성 음성(또는 합성 음성과 연관된 데이터)을 통신 모듈(336) 및 네트워크(220)를 통해 사용자 단말(210)로 제공하고, 생성된 합성 음성을 사용자 단말(210)의 음성 출력 가능 장치 등을 통해 출력하도록 구성될 수 있다. 다른 예로서, 정보 처리 시스템(230)의 프로세서(334)는 통신 모듈(336) 및 네트워크(220)를 통해 사용자 단말(210)로 변환 음성(또는 변환 음성과 연관된 데이터)을 제공하고, 사용자 단말(210)의 음성 출력 가능 장치 등을 통해 출력하도록 구

성될 수 있다.

[0054] 도 4는 본 개시의 일 실시예에 따른 음성 합성 모델을 학습시키는 방법의 예시를 나타내는 도면이다. 음성 합성 모델(예를 들어, 음향 특징 인코더(420), 음성 생성기(422), 음향 특징 제거기(430), 자기지도 학습 모델(440), 언어 특징 인코더(442), 자기지도 표현 제거기(450), 텍스트 인코더(462) 등)은 기계학습 모델(예를 들어, 인공신경망 모델)로 구현될 수 있으며, 학습 음성(410) 및/또는 학습 음성(410)에 대응하는 학습 텍스트(460)에 기초하여 학습될 수 있다. 예를 들어, 음성 합성 모델은 총 손실을 최소화하도록 학습될 수 있으며, 총 손실(total loss)은 제1 손실 내지 제4 손실 중 적어도 하나를 기초로 산출될 수 있다. 예를 들어, 제1 손실 내지 제4 손실은 아래와 같이 산출될 수 있다.

[0055] 정보 처리 시스템은 음향 특징 인코더(420)를 이용하여 학습 음성(410)을 기초로 학습 음향 특징을 생성할 수 있으며, 음성 생성기(422)를 이용하여 음향 특징 인코더(420)에 의해 생성된 학습 음향 특징을 기초로 학습 합성 음성(424)을 생성할 수 있다. 제1 손실은 음성 생성기(422)에 의해 생성된 학습 합성 음성(424) 및 학습 음성(410)을 기초로 산출될 수 있다. 예를 들어, 제1 손실은 학습 합성 음성(424)과 학습 음성(410)(GT)에 기초하여 산출된 복원 손실(reconstruction loss)일 수 있다.

[0056] 정보 처리 시스템은 음향 특징 제거기(430)를 이용하여, 음향 특징 인코더(420)에 의해 생성된 학습 음향 특징으로부터 음향적 특징을 제거함으로써 제1 학습 자기지도 표현을 생성할 수 있다. 또한, 정보 처리 시스템은 자기지도 학습 모델(440) 및 언어 특징 인코더(442)를 이용하여 학습 음성(410)으로부터 제2 학습 자기지도 표현을 생성할 수 있다. 제2 손실은 음향 특징 제거기(430)에 의해 생성된 제1 학습 자기지도 표현과 언어 특징 인코더(442)에 의해 생성된 제2 학습 자기지도 표현을 기초로 산출될 수 있다. 예를 들어, 제2 손실은 제1 학습 자기지도 표현과 제2 학습 자기지도 표현에 기초하여 산출된 KL 손실(KL-Divergence loss)일 수 있다.

[0057] 일 실시예에 따르면, 자기지도 학습 모델(440)은 미리 학습된 음성 모델의 적어도 하나의 레이어(예를 들어, wav2vec 2.0의 middle layer)를 포함할 수 있다. 여기서, 미리 학습된 음성 모델의 적어도 하나의 레이어는 언어적 정보를 많이 포함하는 출력을 생성하는 레이어일 수 있다. 음성 합성 모델의 학습 과정에서, 미리 학습된 음성 모델의 적어도 하나의 레이어는 프리징(freezing)될 수 있다.

[0058] 정보 처리 시스템은 자기지도 표현 제거기(450)를 이용하여, 언어 특징 인코더(442)에 의해 생성된 제2 학습 자기지도 표현으로부터 자기지도 표현을 제거함으로써 제1 학습 텍스트 표현을 생성할 수 있다. 또한, 정보 처리 시스템은 텍스트 인코더(462)를 이용하여 학습 텍스트(460)로부터 제2 학습 텍스트 표현을 생성할 수 있다. 제3 손실은 자기지도 표현 제거기(450)에 의해 생성된 제1 학습 텍스트 표현 및 텍스트 인코더(462)로부터 생성된 제2 학습 텍스트 표현을 기초로 산출될 수 있다. 예를 들어, 제3 손실은 제1 학습 텍스트 표현과 제2 학습 텍스트 표현에 기초하여 산출된 KL 손실일 수 있다.

[0059] 정보 처리 시스템은 텍스트 예측기(444)를 이용하여, 언어 특징 인코더(442)에 의해 생성된 제2 학습 자기지도 표현으로부터 단위 텍스트(446)를 예측할 수 있다. 여기서, 단위 텍스트(446)는 학습 텍스트(460)에 포함되는 것으로 예측되는 단위 텍스트(예를 들어, 음소, 글자, 음절 또는 단어 등)일 수 있다. 제4 손실은 텍스트 예측기(444)에 의해 예측된 단위 텍스트(446) 및 학습 텍스트(460)에 기초하여 산출될 수 있다. 예를 들어, 제4 손실은 학습 텍스트(460)에 포함되는 것으로 예측되는 음소를 예측하기 위한 텍스트 예측기(444)에 의해 예측된 음소와 학습 텍스트(460)에 포함된 음소(GT)에 기초하여 산출된 음소 예측 손실(phoneme prediction loss)일 수 있다.

[0060] 총 손실은 상술한 제1 손실 내지 제4 손실 중 적어도 하나를 기초로 산출될 수 있으며, 구체적 예로, 총 손실은 제1 손실 내지 제4 손실의 가중합으로 산출될 수 있다. 음성 합성 모델은 총 손실을 최소화하도록 학습될 수 있다.

[0061] 도 5는 본 개시의 일 실시예에 따른 학습된 음성 합성 모델을 이용하여 합성 음성(520)을 생성하는 방법의 예시를 나타내는 도면이다. 일 실시예에 따르면, 정보 처리 시스템은 도 4를 참조하여 상술한 바와 같이 학습되고/되거나 도 7을 참조하여 후술하는 바와 같이 학습된 음성 합성 모델을 이용하여, 입력 텍스트(510)에 대한 합성 음성(520)을 생성할 수 있다. 예를 들어, 정보 처리 시스템은 텍스트 인코더(462)를 이용하여, 입력 텍스트(510)로부터 텍스트 표현을 생성하고, 자기지도 표현 생성기(452)를 이용하여, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성할 수 있다. 그런 다음, 정보 처리 시스템은 음향 특징 생성기(432)를 이용하여, 자기지도 표현을 기초로 음향 특징을 생성할 수 있고, 음성 생성기(422)를 이용하여, 음향 특징을 기초로 합성 음성(520)을 생성할 수 있다.

[0062] 일 실시예에 따르면, 음향 특징 제거기(430), 자기지도 표현 제거기(450)는 역변환이 가능한 네트워크(예를 들어, Normalizing Flow, Denoising Diffusion Model 등)일 수 있다. 이 경우, 음향 특징 생성에 이용되는 음향 특징 생성기(432), 자기지도 표현 생성에 이용되는 자기지도 표현 생성기(452)는 학습된 음향 특징 제거기(430), 자기지도 표현 제거기(450)가 각각 역변환된 모델일 수 있다. 예를 들어, 역변환이 가능한 단일 네트워크를 연산 방향만 달리하여 사용함으로써, 음향 특징 제거기(430) 또는 음향 특징 생성기(432)가 사용될 수 있다. 이와 유사하게, 역변환이 가능한 단일 네트워크를 연산 방향만 달리하여 사용함으로써, 자기지도 표현 제거기(450) 또는 자기지도 표현 생성기(452)가 사용될 수 있다.

[0063] 도 6은 본 개시의 일 실시예에 따른 변환 음성(660)을 생성하는 방법의 예시를 나타내는 도면이다. 일 실시예에 따르면, 음성 합성은 소스 음성(610)을 기초로 변환 음성(660)을 생성하는 음성 변환을 포함할 수 있다. 예를 들어, 음성 변환은 소스 음성(610)의 발화 텍스트, 발음 등 언어적 정보는 동일하게 유지하고, 음향 특징(예를 들어, 목소리)만을 변환함으로써 변환 음성(660)을 생성하는 것을 지칭할 수 있다.

[0064] 일 실시예에 따르면, 정보 처리 시스템은 소스 음성(610), 소스 음성 정보(612) 및 타겟 음성 정보(642)를 수신할 수 있다. 예를 들어, 소스 음성 정보(612)는 소스 음성(610)과 별도로 수신되거나, 소스 음성(610)으로부터 추출될 수 있다. 타겟 음성 정보(642)는 별도로 수신되거나, 타겟 음성 샘플로부터 추출될 수 있다.

[0065] 정보 처리 시스템은 음성 변환 모델을 이용하여 소스 음성(610)을 기초로 변환 음성(660)을 생성할 수 있다. 음성 변환 모델은 음향 특징 인코더(620), 음향 특징 제거기(630), 음향 특징 생성기(640) 및 음성 생성기(650)를 포함할 수 있다. 일 실시예에 따르면, 음성 변환 모델은 도 4를 참조하여 상술한 바와 같이 학습되고/되거나 도 7을 참조하여 후술하는 바와 같이 학습될 수 있다.

[0066] 음향 특징 인코더(620)는 소스 음성(610) 및 소스 음성 정보(612)를 입력으로 수신하고, 소스 음성(610) 및 소스 음성 정보(612)에 기초하여 소스 음향 특징을 생성할 수 있다. 예를 들어, 소스 음향 특징은 소스 음성(610)의 음향적 정보를 포함하는 임베딩 벡터일 수 있다.

[0067] 음향 특징 제거기(630)는 음향 특징 인코더(620)에 의해 생성된 소스 음향 특징 및 소스 음성 정보(612)를 입력으로 수신하고, 소스 음향 특징 및 소스 음성 정보(612)를 기초로 자기지도 표현을 생성할 수 있다. 예를 들어, 음향 특징 제거기(630)는 소스 음성 정보(612)를 이용하여, 소스 음향 특징으로부터 음향적 특징이 제거되고 언어적 정보를 포함하는 자기지도 표현을 생성할 수 있다.

[0068] 음향 특징 생성기(640)는 음향 특징 제거기(630)에 의해 생성된 자기지도 표현 및 타겟 음성 정보(642)를 입력으로 수신하고, 자기지도 표현 및 타겟 음성 정보(642)를 기초로 타겟 음향 특징을 생성할 수 있다. 예를 들어, 음향 특징 생성기(640)는 타겟 음성 정보(642)를 이용하여, 자기지도 표현으로부터 타겟 음성의 음향적 특징이 포함된 타겟 음향 특징을 생성할 수 있다. 일 실시예에 따르면, 음향 특징 생성기(640)는 학습된 음향 특징 제거기(630)가 역변환된 모델일 수 있다.

[0069] 음성 생성기(650)는 음향 특징 생성기(640)에 의해 생성된 타겟 음향 특징 및 타겟 음성 정보(642)를 입력으로 수신하고, 타겟 음향 특징 및 타겟 음성 정보를 기초로 변환 음성(660)을 생성할 수 있다.

[0070] 이와 같이, 음성의 언어적 정보를 유지하면서 음향 특징만을 제거하도록 학습된 음향 특징 제거기(630) 및 학습된 음향 특징 제거기(630)가 역변환된 음향 특징 생성기(640)를 이용하여, 소스 음성(610)의 언어적 정보는 그대로 유지하면서 음향 특징만 변환된 변환 음성(660)을 생성할 수 있다.

[0071] 도 7은 본 개시의 일 실시예에 따라 새로운 화자에 대한 학습 음성을 이용하여 음성 합성 모델을 전이 학습(fine-tuning)시키는 방법의 예시를 나타내는 도면이다. 추가적으로, 음성 합성 모델(예를 들어, 음성 생성기(722), 음향 특징 제거기(730) 등)은 새로운 화자에 대한 학습 음성(710)을 기초로 전이 학습될 수 있다. 즉, 새로운 화자에 대한 학습 음성(710)에 대응하는 학습 텍스트 없이, 음성 샘플만으로 음성 합성 모델을 새로운 화자에 적응시킬 수 있다.

[0072] 예를 들어, 음성 합성 모델은 총 손실을 최소화하도록 전이 학습될 수 있으며, 총 손실은 제1 손실 또는 제2 손실 중 적어도 하나를 기초로 산출될 수 있다. 여기서, 제1 손실 또는 제2 손실은 도 4를 참조하여 설명한 바와 동일 또는 유사하게 산출될 수 있다. 예를 들어, 제1 손실, 제2 손실은 아래와 같이 산출될 수 있다.

[0073] 정보 처리 시스템은 음향 특징 인코더(720)를 이용하여 학습 음성(710)을 기초로 학습 음향 특징을 생성할 수 있으며, 음성 생성기(722)를 이용하여 음향 특징 인코더(720)에 의해 생성된 학습 음향 특징을 기초로 학습 합성 음성(724)을 생성할 수 있다. 제1 손실은 음성 생성기(722)에 의해 생성된 학습 합성 음성(724) 및 학습 음

성(710)을 기초로 산출될 수 있다. 예를 들어, 제1 손실은 학습 합성 음성(724)과 학습 음성(710)(GT)에 기초하여 산출된 복원 손실일 수 있다.

[0074] 정보 처리 시스템은 음향 특징 제거기(730)를 이용하여, 음향 특징 인코더(720)에 의해 생성된 학습 음향 특징으로부터 음향적 특징을 제거함으로써 제1 학습 자기지도 표현을 생성할 수 있다. 또한, 정보 처리 시스템은 자기지도 학습 모델(740) 및 언어 특징 인코더(742)를 이용하여 학습 음성(710)으로부터 제2 학습 자기지도 표현을 생성할 수 있다. 제2 손실은 음향 특징 제거기(730)에 의해 생성된 제1 학습 자기지도 표현과 언어 특징 인코더(742)에 의해 생성된 제2 학습 자기지도 표현을 기초로 산출될 수 있다. 예를 들어, 제2 손실은 제1 학습 자기지도 표현과 제2 학습 자기지도 표현에 기초하여 산출된 KL 손실일 수 있다.

[0075] 이와 같이, 새로운 화자에 대한 학습 음성(710)에 대응하는 학습 텍스트 없이, 음성 샘플만으로 음성 합성 모델을 새로운 화자에 적응시킬 수 있다. 정보 처리 시스템은 새로운 화자에 대한 학습 음성(710)에 기초하여 전이 학습된 음성 합성 모델을 이용하여, 입력 텍스트를 기초로 새로운 화자의 음성 특징을 반영하는 합성 음성을 생성할 수 있다.

[0076] 도 8은 본 개시의 일 실시예에 따른 음성 합성 모델의 성능을 평가한 예시를 나타내는 도면이다. 도 8에는 여러가지 성능 평가 지표를 이용하여, 본 개시의 일 실시예에 따른 음성 합성 모델을 포함한 여러가지 음성 합성 모델의 성능을 평가한 결과표가 도시되어 있다.

[0077] 표 1(810)는 VCTK 데이터 세트를 이용한 텍스트-음성 합성 성능 평가 결과를 나타내는 표이다. MOS(Mean Opinion Score)(nMOS(백그라운드 잡음과 관련됨) 및 sMOS(포어그라운드 품질과 관련됨))와 관련한 성능 평가 결과를 살펴보면, 다른 음성 합성 모델에 대비하여 본 방법의 음성 합성 모델(This Method로 도시됨)의 성능이 가장 우수한 것이 확인될 수 있다. 또한, ASR(Automatic Speech Recognition) 평가 측면에서, 본 방법의 음성 합성 모델은 보다 정확한 발음으로 음성을 합성하여 다른 모델들에 대비하여 가장 낮은 PER(Phoneme Error Rate)과 WER(Word Error Rate)을 보인 것이 확인될 수 있다. 또한, MCD(Mel-Cepstral Distortion) 및 RMSE(Root Mean Squared Error) 측면에서 본 방법의 음성 합성 모델은 가장 낮은 오차 거리를 가지는 것이 확인될 수 있다.

[0078] 표 2(820)는 본 방법의 음성 합성 모델을 새로운 화자에 대한 학습 음성 및 학습 음성에 대응하는 학습 텍스트에 기초하여 전이 학습시킨 모델(Method 1로 도시됨, 이하 제1 모델) 및 학습 텍스트 없이 새로운 화자에 대한 학습 음성에만 기초하여 전이 학습시킨 모델(Method 2로 도시됨, 이하 제2 모델)의 텍스트-음성 합성 성능을 평가한 결과를 나타내는 표이다. MOS(nMOS 및 sMOS) 평가 측면에서 제2 모델은 제1 모델과 유사한 성능을 가짐이 확인될 수 있다. 또한, 제2 모델은 제1 모델보다 낮은 PER 및 WER을 보인 것이 확인될 수 있다. 이는, 학습 텍스트 없이 새로운 화자에 대한 학습 음성만으로 본 방법의 모델을 전이 학습시키더라도, 모델을 새로운 화자에 잘 적응시킬 수 있음을 보여준다.

[0079] 표 3(830)는 VCTK 데이터 세트를 이용한 음성 변환(Voice Conversion) 성능 평가 결과를 나타내는 표이다. 본 방법의 음성 변환 모델(This Method로 도시됨)은 nMOS 및 ASR 평가 지표에 있어서, 다른 모델들에 대비하여 가장 성능이 우수한 것이 확인될 수 있다. 이는, 본 방법의 음성 변환 모델이 언어적 정보의 최대한 작은 손실로 언어적 정보와 음향적 정보를 분리할 수 있음을 의미할 수 있다.

[0080] 도 9는 본 개시의 일 실시예에 따른 음성 합성 방법(900)의 예시를 나타내는 흐름도이다. 일 실시예에 따르면, 방법(900)은 프로세서(예를 들어, 정보 처리 시스템의 적어도 하나의 프로세서 및/또는 사용자 단말의 적어도 하나의 프로세서)가 입력 텍스트를 수신함으로써 개시될 수 있다(S910). 음성 합성 모델이 다화자 모델인 경우, 프로세서는 타겟 음성 정보를 추가적으로 수신할 수 있다.

[0081] 그런 다음, 프로세서는 텍스트 인코더를 이용하여, 입력 텍스트로부터 텍스트 표현을 생성하고(S920), 자기지도 표현 생성기를 이용하여, 텍스트 표현으로부터 언어적 정보를 포함하는 자기지도 표현을 생성할 수 있다(S930). 일 실시예에서 자기지도 표현 생성기는, 학습된 자기지도 표현 제거기가 역변환된 모델일 수 있다.

[0082] 그런 다음, 프로세서는 음향 특징 생성기를 이용하여, 자기지도 표현을 기초로 음향 특징을 생성할 수 있다(S940). 추가적으로 또는 대안적으로, 음성 합성 모델이 다화자 모델인 경우, 프로세서는 음향 특징 생성기를 이용하여, 자기지도 표현 및 타겟 음성 정보를 기초로 음향 특징을 생성할 수 있다. 일 실시예에 따르면, 음향 특징 생성기는 학습된 음향 특징 제거기가 역변환된 모델일 수 있다.

[0083] 그 후, 프로세서는 음성 생성기를 이용하여, 음향 특징을 기초로 합성 음성을 생성할 수 있다(S950). 추가적으로 또는 대안적으로, 음성 합성 모델이 다화자 모델인 경우, 프로세서는 음성 생성기를 이용하여, 음향 특징 및

타겟 음성 정보를 기초로 합성 음성을 생성할 수 있다.

[0084] 일 실시예에 따르면, 음성 합성 모델(예를 들어, 자기지도 표현 제거기, 음향 특징 제거기 또는 음성 생성기 중 적어도 하나)은 제1 학습 음성 또는 제1 학습 음성에 대응하는 제1 학습 텍스트 중 적어도 하나에 기초하여 학습될 수 있다. 예를 들어, 음성 합성 모델은 총 손실을 최소화하도록 학습될 수 있으며, 총 손실은 제1 손실 내지 제4 손실 중 적어도 하나를 기초로 산출될 수 있다.

[0085] 구체적 예로, 제1 손실은 학습 합성 음성 및 제1 학습 음성에 기초하여 산출될 수 있고, 학습 합성 음성은 음향 특징 인코더 및 음성 생성기를 이용하여 제1 학습 음성을 기초로 생성될 수 있다. 또한, 제2 손실은 제1 학습 자기지도 표현 및 제2 학습 자기지도 표현에 기초하여 산출될 수 있다. 여기서, 제1 학습 자기지도 표현은 음향 특징 인코더 및 음향 특징 제거기를 이용하여, 제1 학습 음성을 기초로 생성될 수 있으며, 제2 학습 자기지도 표현은 자기지도 학습 모델 및 언어 특징 인코더를 이용하여 제1 학습 음성으로부터 생성될 수 있다.

[0086] 제3 손실은 제1 학습 텍스트 표현 및 제2 학습 텍스트 표현에 기초하여 산출될 수 있다. 여기서, 제1 학습 텍스트 표현은 자기지도 학습 모델, 언어 특징 인코더 및 자기지도 표현 제거기를 이용하여 제1 학습 음성으로부터 생성될 수 있으며, 제2 학습 텍스트 표현은 텍스트 인코더를 이용하여, 제1 학습 텍스트로부터 생성될 수 있다. 또한, 제4 손실은 예측된 단위 텍스트 및 제1 학습 텍스트에 기초하여 산출될 수 있고, 예측된 단위 텍스트는 자기지도 학습 모델, 언어 특징 인코더 및 텍스트 예측기를 이용하여, 제1 학습 음성으로부터 예측될 수 있다. 일 실시예에서, 텍스트 예측기에 의해 예측된 단위 텍스트는 제1 학습 텍스트에 포함되는 것으로 예측되는 음소, 글자, 음절 또는 단어를 포함할 수 있다.

[0087] 일 실시예에 따르면, 자기지도 학습 모델은 미리 학습된 음성 모델의 적어도 하나의 레이어를 포함할 수 있고, 미리 학습된 음성 모델의 적어도 하나의 레이어는, 학습 과정에서 프리징될 수 있다.

[0088] 추가적으로, 음성 합성 모델(예를 들어, 음성 생성기 및 음향 특징 제거기)는 새로운 화자에 대한 제2 학습 음성을 기초로 전이 학습될 수 있다. 이 경우, 프로세서는 제2 학습 음성에 기초하여 전이 학습된 음향 특징 제거기가 역변환된 음향 특징 생성기를 이용하여, 자기지도 표현을 기초로 음향 특징을 생성할 수 있다. 그런 다음, 프로세서는 제2 학습 음성에 기초하여 전이 학습된 음성 생성기를 이용하여, 음향 특징을 기초로 새로운 화자의 음성 특징을 반영하는 합성 음성을 생성할 수 있다.

[0089] 도 10은 본 개시의 일 실시예에 따른 음성 합성 방법(1000)의 예시를 나타내는 흐름도이다. 방법(1000)은 프로세서(예를 들어, 정보 처리 시스템의 적어도 하나의 프로세서 및/또는 사용자 단말의 적어도 하나의 프로세서)가 소스 음성, 소스 음성 정보 및 타겟 음성 정보를 수신함으로써 개시될 수 있다(S1010).

[0090] 그런 다음, 프로세서는 음향 특징 인코더를 이용하여, 소스 음성 및 소스 음성 정보를 기초로 소스 음향 특징을 생성하고(S1020), 음향 특징 제거기를 이용하여, 소스 음향 특징 및 소스 음성 정보를 기초로 언어적 정보를 포함하는 자기지도 표현을 생성할 수 있다(S1030).

[0091] 그런 다음, 프로세서는 음향 특징 생성기를 이용하여, 자기지도 표현 및 타겟 음성 정보를 기초로 타겟 음향 특징을 생성할 수 있다(S1040). 일 실시예에 따르면, 음향 특징 생성기는 학습된 음향 특징 제거기가 역변환된 모델일 수 있다. 그 후, 프로세서는 음성 생성기를 이용하여, 타겟 음향 특징 및 타겟 음성 정보를 기초로 변환 음성을 생성할 수 있다(S1050).

[0092] 일 실시예에 따르면, 음성 변환 모델(예를 들어, 음향 특징 제거기 또는 음성 생성기 중 적어도 하나)은 제3 학습 음성에 기초하여 학습될 수 있다. 예를 들어, 음성 변환 모델은 총 손실을 최소화하도록 학습될 수 있으며, 총 손실은 제5 손실 또는 제6 손실 중 적어도 하나를 기초로 산출될 수 있다.

[0093] 구체적 예로, 제5 손실은 학습 합성 음성 및 제3 학습 음성에 기초하여 산출될 수 있고, 학습 합성 음성은 음향 특징 인코더 및 음성 생성기를 이용하여, 제3 학습 음성을 기초로 생성될 수 있다. 또한, 제6 손실은 제1 학습 자기지도 표현 및 제2 학습 자기지도 표현에 기초하여 산출될 수 있다. 여기서, 제1 학습 자기지도 표현은 음향 특징 인코더 및 음향 특징 제거기를 이용하여, 제3 학습 음성을 기초로 생성될 수 있으며, 제2 학습 자기지도 표현은 자기지도 학습 모델 및 언어 특징 인코더를 이용하여 제3 학습 음성으로부터 생성될 수 있다.

[0094] 일 실시예에 따르면, 자기지도 학습 모델은 미리 학습된 음성 모델의 적어도 하나의 레이어를 포함할 수 있다. 여기서, 미리 학습된 음성 모델의 적어도 하나의 레이어는 학습 과정에서 프리징될 수 있다.

[0095] 도 9, 도 10 및 상술한 설명은 본 개시의 일 실시예일 뿐이며, 본 개시의 범위가 이에 한정되는 것은 아니다. 예를 들어, 다른 실시예에 따르면, 상술한 방법 중 적어도 일부의 단계가 다른 수행 주체에 의해 수행되거나,

각 단계의 순서가 바뀌거나, 적어도 일부 단계가 추가/변경/생략될 수 있다.

[0096] 상술한 방법은 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램으로 제공될 수 있다. 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일수도 있다. 또한, 매체는 단일 또는 수개 하드웨어가 결합된 형태의 다양한 기록 수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 애플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.

[0097] 본 개시의 방법, 동작 또는 기법들은 다양한 수단에 의해 구현될 수도 있다. 예를 들어, 이러한 기법들은 하드웨어, 펌웨어, 소프트웨어, 또는 이들의 조합으로 구현될 수도 있다. 본원의 개시와 연계하여 설명된 다양한 예시적인 논리적 블록들, 모듈들, 회로들, 및 알고리즘 단계들은 전자 하드웨어, 컴퓨터 소프트웨어, 또는 양자의 조합들로 구현될 수도 있음을 통상의 기술자들은 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 대체를 명확하게 설명하기 위해, 다양한 예시적인 구성요소들, 블록들, 모듈들, 회로들, 및 단계들이 그들의 기능적 관점에서 일반적으로 위에서 설명되었다. 그러한 기능이 하드웨어로서 구현되는지 또는 소프트웨어로서 구현되는 지의 여부는, 특정 애플리케이션 및 전체 시스템에 부과되는 설계 요구사항들에 따라 달라진다. 통상의 기술자들은 각각의 특정 애플리케이션을 위해 다양한 방식들로 설명된 기능을 구현할 수도 있으나, 그러한 구현들은 본 개시의 범위로부터 벗어나게 하는 것으로 해석되어서는 안된다.

[0098] 하드웨어 구현에서, 기법들을 수행하는 데 이용되는 프로세싱 유닛들은, 하나 이상의 ASIC들, DSP들, 디지털 신호 프로세싱 디바이스들(digital signal processing devices; DSPD들), 프로그램가능 논리 디바이스들(programmable logic devices; PLD들), 필드 프로그램가능 게이트 어레이들(field programmable gate arrays; FPGA들), 프로세서들, 제어기들, 마이크로제어기들, 마이크로프로세서들, 전자 디바이스들, 본 개시에 설명된 기능들을 수행하도록 설계된 다른 전자 유닛들, 컴퓨터, 또는 이들의 조합 내에서 구현될 수도 있다.

[0099] 따라서, 본 개시와 연계하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 및 회로들은 범용 프로세서, DSP, ASIC, FPGA나 다른 프로그램 가능 논리 디바이스, 이산 게이트나 트랜지스터 로직, 이산 하드웨어 컴포넌트들, 또는 본원에 설명된 기능들을 수행하도록 설계된 것들의 임의의 조합으로 구현되거나 수행될 수도 있다. 범용 프로세서는 마이크로프로세서일 수도 있지만, 대안으로, 프로세서는 임의의 종래의 프로세서, 제어기, 마이크로제어기, 또는 상태 머신일 수도 있다. 프로세서는 또한, 컴퓨팅 디바이스들의 조합, 예를 들면, DSP와 마이크로프로세서, 복수의 마이크로프로세서들, DSP 코어와 연계한 하나 이상의 마이크로프로세서들, 또는 임의의 다른 구성의 조합으로서 구현될 수도 있다.

[0100] 펌웨어 및/또는 소프트웨어 구현에 있어서, 기법들은 랜덤 액세스 메모리(random access memory; RAM), 판독 전용 메모리(read-only memory; ROM), 비휘발성 RAM(non-volatile random access memory; NVRAM), PROM(programmable read-only memory), EPROM(erasable programmable read-only memory), EEPROM(electrically erasable PROM), 플래시 메모리, 컴팩트 디스크(compact disc; CD), 자기 또는 광학 데이터 스토리지 디바이스 등과 같은 컴퓨터 판독가능 매체 상에 저장된 명령들로서 구현될 수도 있다. 명령들은 하나 이상의 프로세서들에 의해 실행 가능할 수도 있고, 프로세서(들)로 하여금 본 개시에 설명된 기능의 특정 양태들을 수행하게 할 수도 있다.

[0101] 소프트웨어로 구현되는 경우, 상기 기법들은 하나 이상의 명령들 또는 코드로서 컴퓨터 판독 가능한 매체 상에 저장되거나 또는 컴퓨터 판독 가능한 매체를 통해 전송될 수도 있다. 컴퓨터 판독가능 매체들은 한 장소에서 다른 장소로 컴퓨터 프로그램의 전송을 용이하게 하는 임의의 매체를 포함하여 컴퓨터 저장 매체들 및 통신 매체들 양자를 포함한다. 저장 매체들은 컴퓨터에 의해 액세스될 수 있는 임의의 이용 가능한 매체들일 수도 있다. 비제한적인 예로서, 이러한 컴퓨터 판독가능 매체는 RAM, ROM, EEPROM, CD-ROM 또는 다른 광학 디스크 스토리지, 자기 디스크 스토리지 또는 다른 자기 스토리지 디바이스들, 또는 소망의 프로그램 코드를 명령들 또는 데이터 구조들의 형태로 이송 또는 저장하기 위해 사용될 수 있으며 컴퓨터에 의해 액세스될 수 있는 임의의 다른 매체를 포함할 수 있다. 또한, 임의의 접속이 컴퓨터 판독가능 매체로 적절히 칭해진다.

[0102] 예를 들어, 소프트웨어가 동축 케이블, 광섬유 케이블, 연선, 디지털 가입자 회선 (DSL), 또는 적외선, 무선, 및 마이크로파와 같은 무선 기술들을 사용하여 웹사이트, 서버, 또는 다른 원격 소스로부터 전송되면, 동축 케

이블, 광섬유 케이블, 연선, 디지털 가입자 회선, 또는 적외선, 무선, 및 마이크로파와 같은 무선 기술들은 매체의 정의 내에 포함된다. 본원에서 사용된 디스크(disk) 와 디스크(disc)는, CD, 레이저 디스크, 광 디스크, DVD(digital versatile disc), 플로피디스크, 및 블루레이 디스크를 포함하며, 여기서 디스크들(disks)은 보통 자기적으로 데이터를 재생하고, 반면 디스크들(discs) 은 레이저를 이용하여 광학적으로 데이터를 재생한다. 위의 조합들도 컴퓨터 판독가능 매체들의 범위 내에 포함되어야 한다.

[0103] 소프트웨어 모듈은, RAM 메모리, 플래시 메모리, ROM 메모리, EPROM 메모리, EEPROM 메모리, 레지스터들, 하드 디스크, 이동식 디스크, CD-ROM, 또는 공지된 임의의 다른 형태의 저장 매체 내에 상주할 수도 있다. 예시적인 저장 매체는, 프로세가 저장 매체로부터 정보를 판독하거나 저장 매체에 정보를 기록할 수 있도록, 프로세서에 연결될 수 있다. 대안으로, 저장 매체는 프로세서에 통합될 수도 있다. 프로세서와 저장 매체는 ASIC 내에 존재할 수도 있다. ASIC은 유저 단말 내에 존재할 수도 있다. 대안으로, 프로세서와 저장 매체는 유저 단말에서 개별 구성요소들로서 존재할 수도 있다.

[0104] 이상 설명된 실시예들이 하나 이상의 독립형 컴퓨터 시스템에서 현재 개시된 주제의 양태들을 활용하는 것으로 기술되었으나, 본 개시는 이에 한정되지 않고, 네트워크나 분산 컴퓨팅 환경과 같은 임의의 컴퓨팅 환경과 연계하여 구현될 수도 있다. 또 나아가, 본 개시에서 주제의 양상들은 복수의 프로세싱 칩들이나 장치들에서 구현될 수도 있고, 스토리지는 복수의 장치들에 걸쳐 유사하게 영향을 받게 될 수도 있다. 이러한 장치들은 PC들, 네트워크 서버들, 및 휴대용 장치들을 포함할 수도 있다.

[0105] 본 명세서에서는 본 개시가 일부 실시예들과 관련하여 설명되었지만, 본 개시의 발명이 속하는 기술분야의 통상의 기술자가 이해할 수 있는 본 개시의 범위를 벗어나지 않는 범위에서 다양한 변형 및 변경이 이루어질 수 있다. 또한, 그러한 변형 및 변경은 본 명세서에 첨부된 특허청구의 범위 내에 속하는 것으로 생각되어야 한다.

부호의 설명

[0106] 110: 입력 텍스트 120: 텍스트 인코더

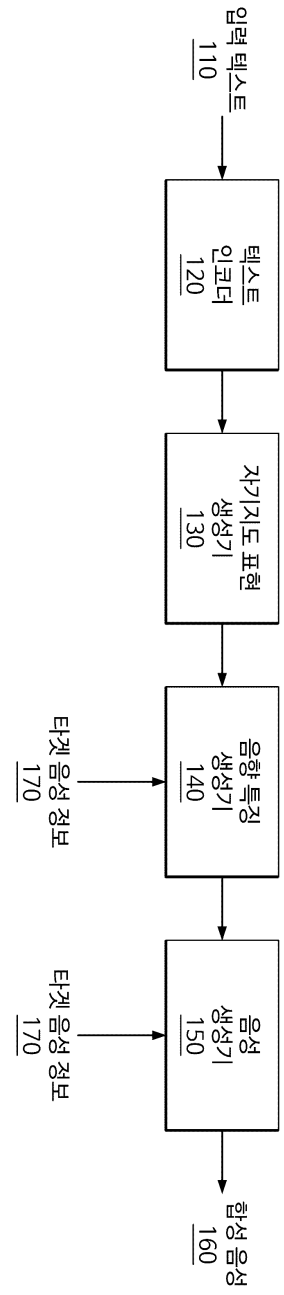
130: 자기지도 표현 생성기 140: 음향 특징 생성기

150: 음성 생성기 160: 합성 음성

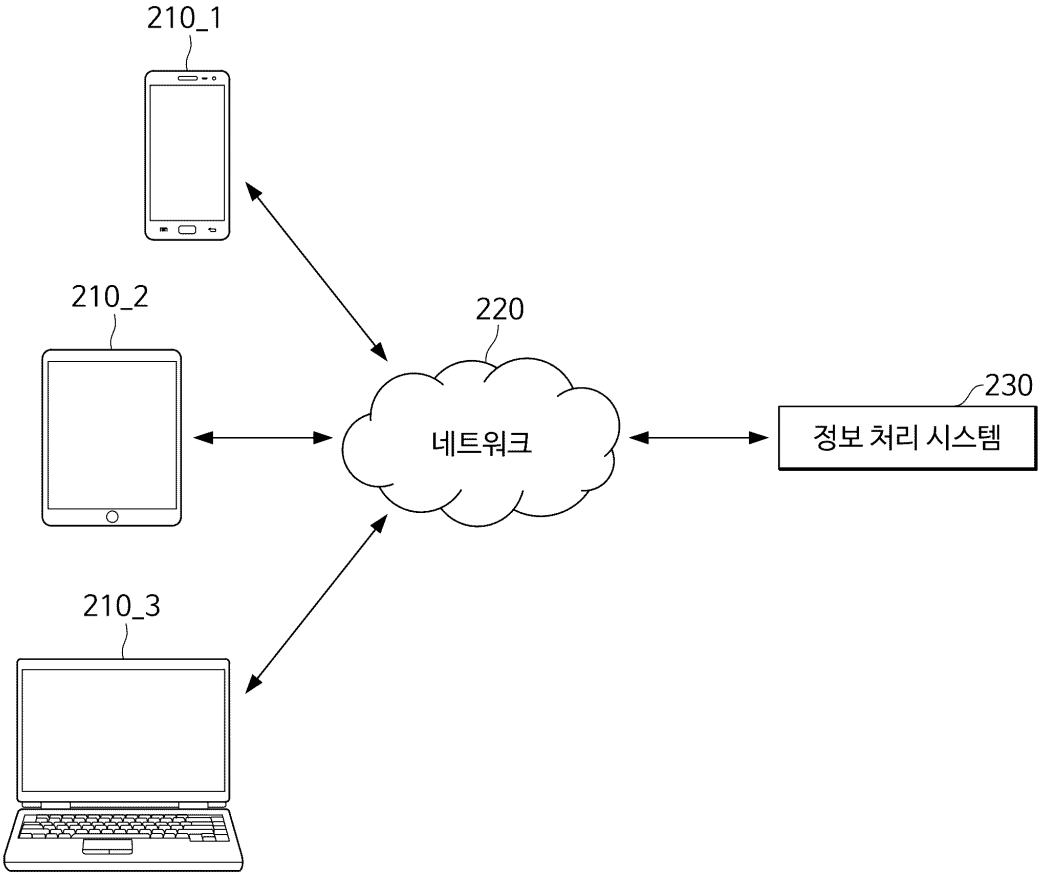
170: 타겟 음성 정보

도면

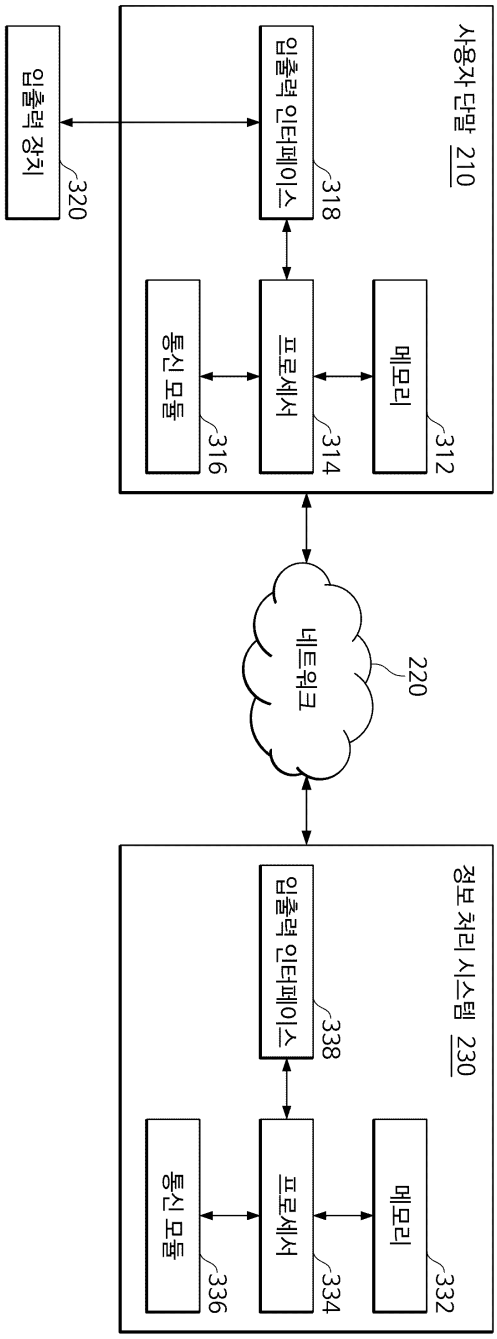
도면1


입력 텍스트
110
텍스트 인코더
120
자기지도 표현 생성기
130
음향 특징 생성기
140
타겟 음성 정보
170
음성 생성기
150
타겟 음성 정보
170
합성 음성
160

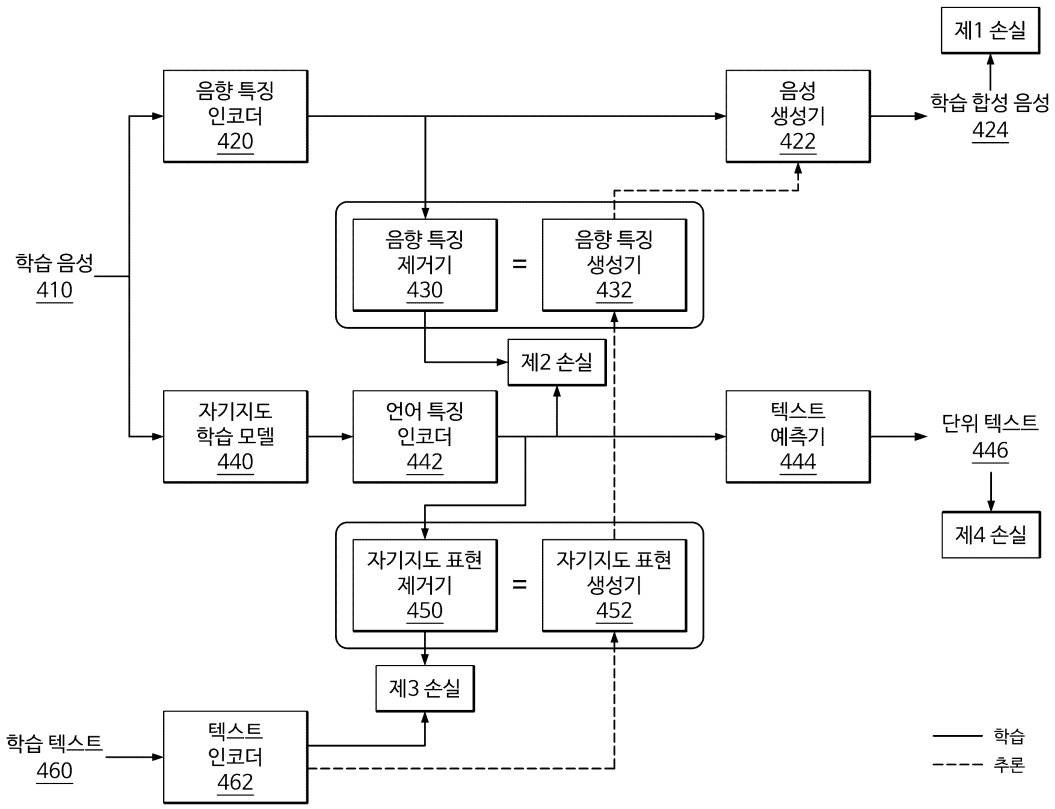
도면2



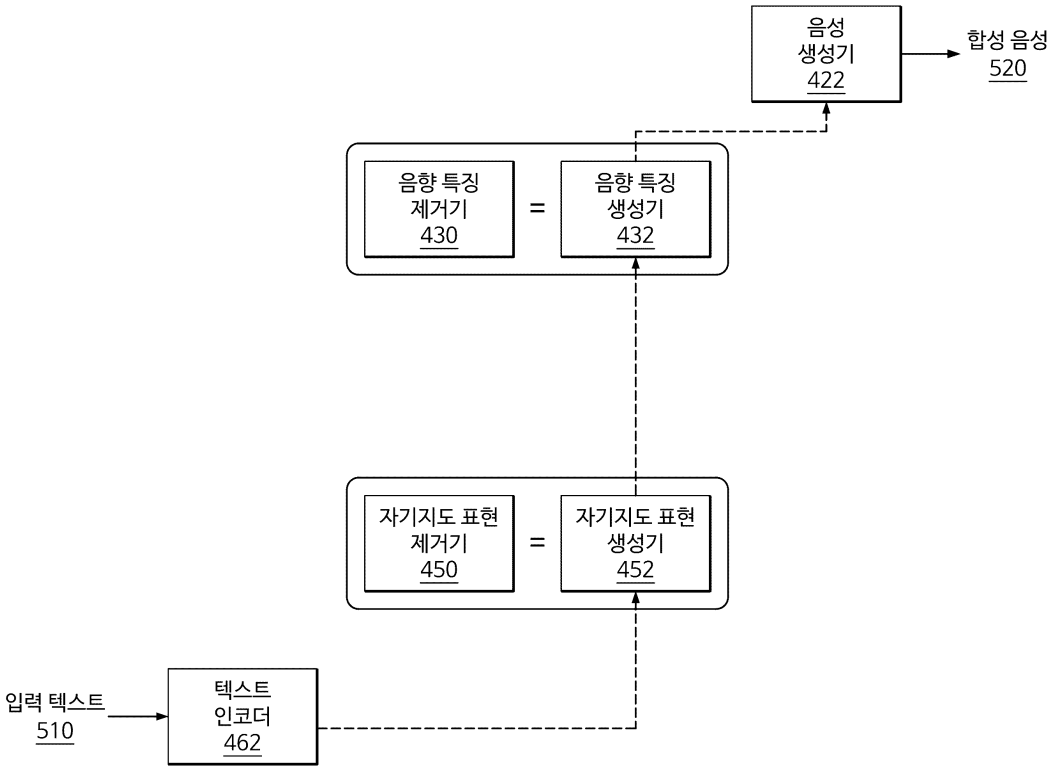
도면3


사용자 단말 210
입출력 인터페이스
318
입출력 장치
320
프로세서
314
메모리
312
통신 모듈
316
네트워크
220
정보 처리 시스템 230
입출력 인터페이스
338
프로세서
334
메모리
332
통신 모듈
336

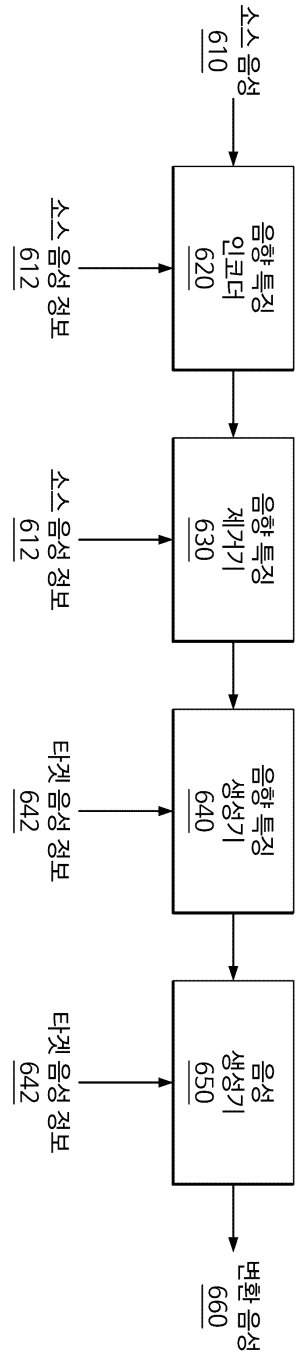
도면4


제1 손실
음향 특징
인코더
420
음성
생성기
422
학습 합성 음성
424
음향 특징
제거기
430
=
음향 특징
생성기
432
학습 음성
410
제2 손실
자기지도
학습 모델
440
언어 특징
인코더
442
텍스트
예측기
444
단위 텍스트
446
제4 손실
자기지도 표현
제거기
450
=
자기지도 표현
생성기
452
제3 손실
학습 텍스트
460
텍스트
인코더
462
학습
추론

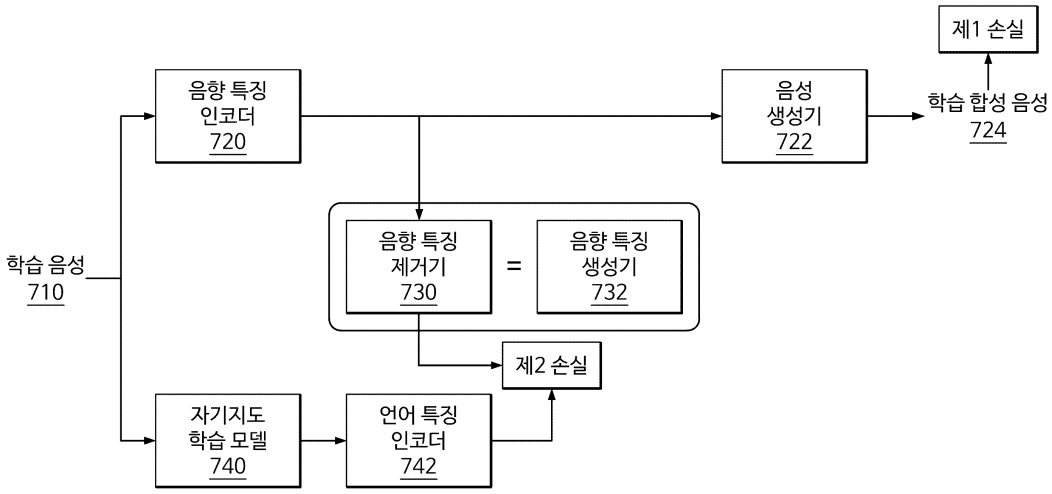
도면5


음성
생성기
422
합성 음성
520
음향 특징
제거기
430
=
음향 특징
생성기
432
자기지도 표현
제거기
450
=
자기지도 표현
생성기
452
입력 텍스트
510
텍스트
인코더
462

도면6


소스 음성
610
음향 특징
인코더
620
소스 음성 정보
612
음향 특징
제거기
630
소스 음성 정보
612
음향 특징
생성기
640
타겟 음성 정보
642
음성
생성기
650
타겟 음성 정보
642
변환 음성
660

도면7



도면8

Method	nMOS	sMOS	PER	WER	EER	MCD	RMSE$_{f0}$	DDUR	Speed (kHz)	Real-time
GT	4.06±0.02	3.34±0.03	5.64	18.94	4.03	-	-	-	-	-
GT (HiFi-GAN)	4.03±0.02	3.30±0.03	5.94	19.52	5.04	1.25	28.32	-	6,484.09	×294.06
Tacotron2	3.76±0.02	3.16±0.03	11.73	22.48	9.11	4.18	35.30	0.49	263.94	×11.97
Glow-TTS	3.95±0.02	3.09±0.03	11.77	26.40	5.33	4.31	32.98	0.38	1,410.75	×63.97
PortaSpeech	3.97±0.02	3.15±0.03	11.35	25.46	5.48	4.34	32.89	0.43	1,163.21	×52.75
VITS	4.02±0.02	3.19±0.03	9.16	25.54	3.83	4.27	32.93	0.37	**1,610.77**	×**72.83**
This Method	**4.04±0.02**	**3.22±0.03**	**5.78**	**19.55**	**3.74**	**4.05**	**32.15**	**0.33**	1,459.95	×66.21

810

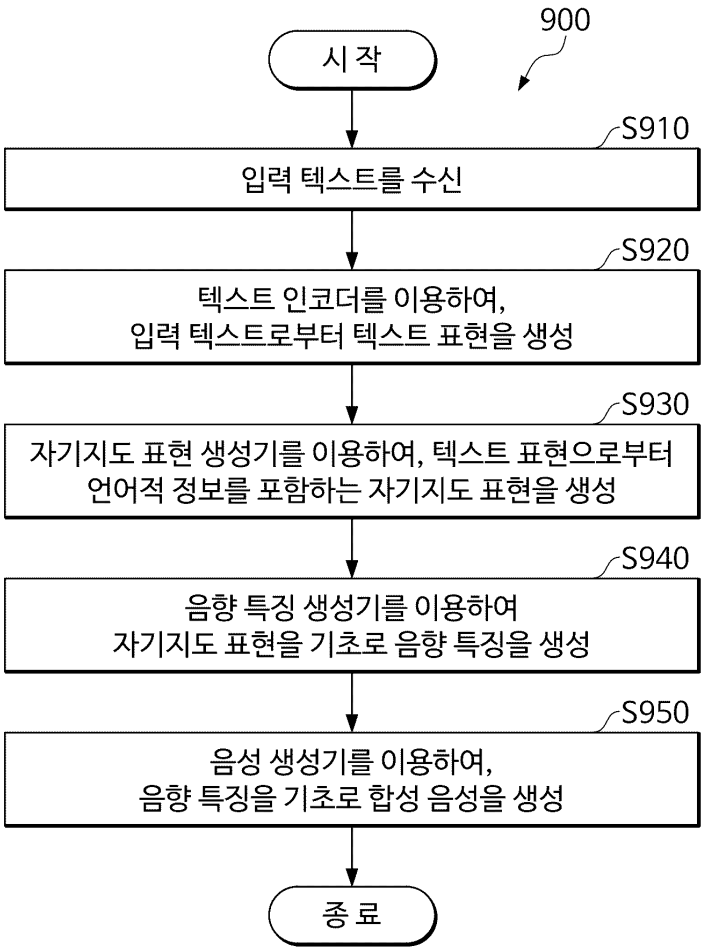
Method	Transcript	nMOS	sMOS	PER	WER	EER	MCD	RMSE$_{f0}$	DDUR
GT	-	4.13±0.10	3.38±0.10	4.26	16.69	4.14	-	-	-
Method 1	✓	4.09±0.10	3.18±0.11	4.40	16.95	6.40	3.96	29.56	0.28
Method 2	✗	4.08±0.09	3.15±0.12	3.71	15.85	6.40	4.09	30.64	0.36

820

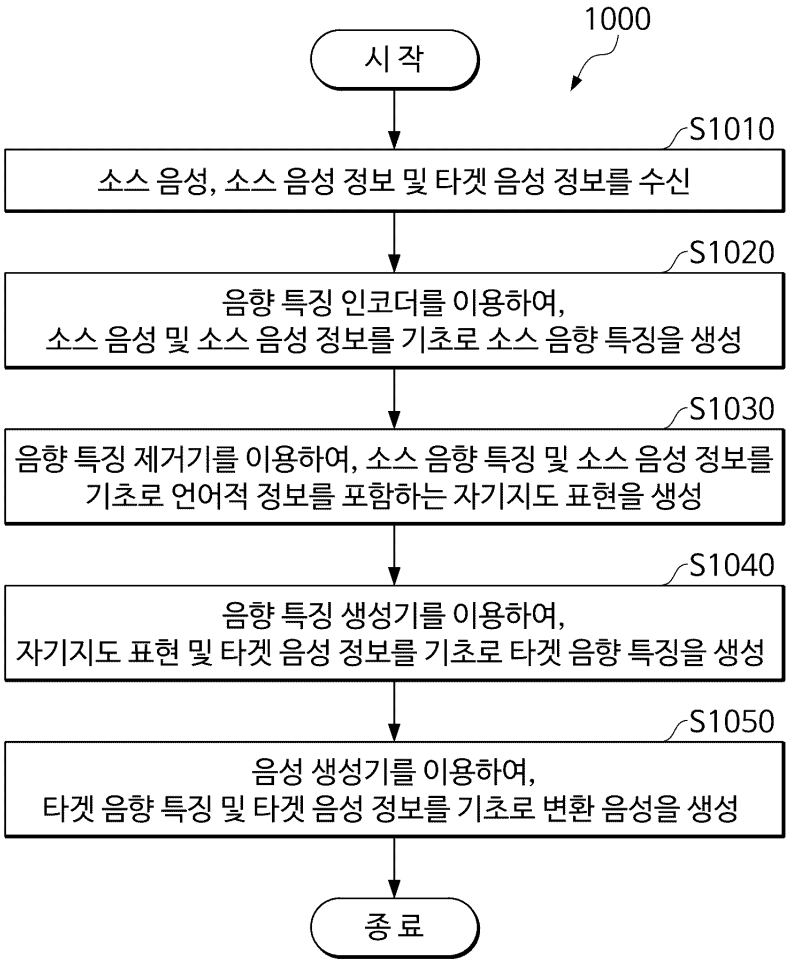
Model	nMOS	sMOS	PER	WER	EER
GT	4.19±0.03	3.37±0.04	5.20	14.05	0.98
AutoVC	3.74±0.05	3.04±0.05	59.42	89.01	12.50
VoiceMixer	4.00±0.05	3.11±0.05	12.67	31.80	9.35
VITS	4.06±0.04	3.16±0.05	6.46	18.73	**2.75**
This Method	**4.08±0.04**	3.16±0.05	**5.51**	**16.25**	**2.75**

830

도면9


900
시 작
S910
입력 텍스트를 수신
S920
텍스트 인코더를 이용하여,
입력 텍스트로부터 텍스트 표현을 생성
S930
자기지도 표현 생성기를 이용하여, 텍스트 표현으로부터
언어적 정보를 포함하는 자기지도 표현을 생성
S940
음향 특징 생성기를 이용하여
자기지도 표현을 기초로 음향 특징을 생성
S950
음성 생성기를 이용하여,
음향 특징을 기초로 합성 음성을 생성
종 료

도면10


1000
시 작
S1010
소스 음성, 소스 음성 정보 및 타겟 음성 정보를 수신
S1020
음향 특징 인코더를 이용하여,
소스 음성 및 소스 음성 정보를 기초로 소스 음향 특징을 생성
S1030
음향 특징 제거기를 이용하여, 소스 음향 특징 및 소스 음성 정보를
기초로 언어적 정보를 포함하는 자기지도 표현을 생성
S1040
음향 특징 생성기를 이용하여,
자기지도 표현 및 타겟 음성 정보를 기초로 타겟 음향 특징을 생성
S1050
음성 생성기를 이용하여,
타겟 음향 특징 및 타겟 음성 정보를 기초로 변환 음성을 생성
종 료