(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0178919
(43) 공개일자 2024년12월31일

(51) 국제특허분류(Int. Cl.)
G10L 13/08 (2006.01) *G10L 13/02* (2006.01)
G10L 17/02 (2013.01) *G10L 19/038* (2013.01)
G10L 25/18 (2013.01) *G10L 25/63* (2013.01)
(52) CPC특허분류
G10L 13/08 (2013.01)
G10L 13/02 (2013.01)
(21) 출원번호 10-2023-0081314
(22) 출원일자 2023년06월23일
심사청구일자 2023년06월23일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
구명완
서울특시 양천구 목동동로 100, 목동아파트. 1312동 1304호
최예린
서울특별시 마포구 대흥로 104 세현오피스텔 603호
(74) 대리인
이지연

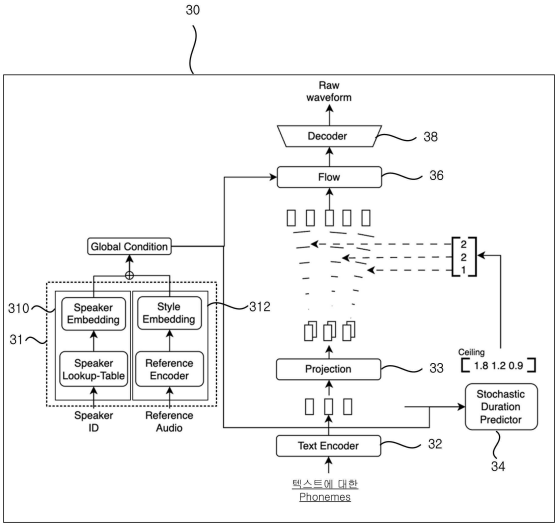
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **양자화된 벡터를 이용한 감정 음성 합성 시스템, 감정 음성 합성 장치 및 모델 학습 장치**

(57) 요 약

본 발명은 양자화된 벡터를 이용한 감정 음성 합성 시스템에 관한 것이다. 상기 감정 음성 합성 시스템은, 텍스트와 정답 음성의 쌍으로 이루어진 학습 데이터 및 화자 룩업 테이블을 이용하여 감정 음성 합성 모델을 학습시키는 모델 학습 장치; 및 학습된 감정 음성 합성 모델을 이용하여, 입력 텍스트를 특정 화자의 감정 음성으로 합성하여 출력하는 감정 음성 합성 장치;를 구비한다. 상기 모델 학습 장치 및 상기 감정 음성 합성 장치는, 텍스트에 대한 정답 음성 또는 참조 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈; 및 화자 룩업 테이블로부터 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈;을 구비하고, 상기 운율 임베딩 벡터 및 상기 화자 임베딩 벡터를 감정 음성 합성 모델에 대한 전역 변수로 제공하는 것을 특징으로 한다.

대 표 도 - 도5



(52) CPC특허분류
G10L 17/02 (2013.01)
G10L 19/038 (2013.01)
G10L 25/18 (2013.01)
G10L 25/63 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1375027496
과제번호	R2022060001-0001
부처명	문화체육관광부
과제관리(전문)기관명	한국콘텐츠진흥원
연구사업명	CT기반조성
연구과제명	인공지능 기반 어린이 독서활동 지원 로봇 및 서비스콘텐츠 개발
기 여 율	1/1
과제수행기관명	광주과학기술원
연구기간	2022.04.01 ~ 2024.12.31

명 세 서

청구범위

청구항 1

텍스트와 이에 대한 정답 음성의 쌍으로 이루어진 데이터 셋을 포함하는 학습 데이터 및 각 화자 정보에 대응되는 화자 임베딩 벡터가 저장된 화자 룩업 테이블을 구비하는 데이터 저장 장치; 학습 데이터 및 화자 룩업 테이블을 이용하여 감정 음성 합성 모델을 학습시키는 모델 학습 장치; 및 학습된 감정 음성 합성 모델을 이용하여, 입력 텍스트를 특정 화자의 감정 음성으로 합성하여 출력하는 감정 음성 합성 장치;를 구비하는 감정 음성 합성 시스템에 있어서,

상기 모델 학습 장치 및 상기 감정 음성 합성 장치는,

음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈; 및

화자 룩업 테이블로부터 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈;

을 구비하고, 상기 운율 임베딩 벡터 및 상기 화자 임베딩 벡터를 감정 음성 합성 모델에 대한 전역 변수로 제공하는 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 2

제1항에 있어서, 상기 모델 학습 장치는,

텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈;

상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈;

텍스트에 대응되는 음소 시퀀스가 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈;

음소 길이 예측 모듈을 위한 정답을 만들기 위하여, 텍스트와 정답 음성에 대한 정렬 벡터를 추정하는 정렬 검색 모듈;

텍스트에 대한 정답 음성에 대한 Linear Spectrum을 이용하여, 정답 음성의 음성 정보를 포함하는 음성 정답 벡터를 생성하여 출력하는 사후 인코더 모듈;

역연산이 가능한 플로우 모델을 구비하고, 상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터에 대하여 정답 음성의 길이를 갖는 텍스트 정보 벡터를 출력하도록 상기 플로우 모델을 학습시키는 플로우 모듈; 및

상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;

을 구비하는 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 3

제2항에 있어서, 상기 모델 학습 장치는

상기 운율 임베딩 벡터 및 화자 임베딩 벡터를 상기 플로우 모듈, 정렬 검색 모듈, 음소 길이 예측 모듈, 사후 인코더 모듈 및 디코더 모듈로 제공하는 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 4

제1항에 있어서, 상기 운율 임베딩 검출 모듈은,

순차적으로 연결된 인코더, 양자화 모듈 및 디코더로 이루어진 오디오 코덱; 및

상기 오디오 코덱의 양자화 모듈로부터 출력된 양자화된 벡터가 입력되며, 상기 입력된 양자화된 벡터로부터 운

율 정보를 포함하는 운율 임베딩 벡터를 추출하여 출력하는 운율 인코더 모듈;을 구비하고,

상기 운율 임베딩 검출 모듈은 참조 음성 또는 정답 음성이 입력되고, 오디오 코덱의 양자화 모듈로부터 상기 입력된 음성에 대한 양자화된 벡터가 출력되고, 상기 양자화된 벡터는 운율 인코더 모듈로 입력되어 운율 임베딩 벡터가 추출되도록 구성된 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 5

제4항에 있어서, 상기 운율 인코더 모듈은,

상기 양자화된 벡터에 대하여 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞는 Time-mixing block과 Channel mixing block을 구비하는 Mixer_TTS 구조; 및

상기 Mixer_TTS 구조에서 구한 특징값을 문장 단위의 운율 임베딩 벡터로 변환시켜 출력하는 GRU(Gated Recurrent Unit);

을 구비하여, 운율 임베딩 벡터를 생성하여 출력하는 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 6

제1항에 있어서, 상기 감정 음성 합성 장치는,

입력 텍스트, 참조 음성 및 화자 정보를 입력받는 입력 모듈;

입력 텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈;

상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈;

입력 텍스트에 대한 음소 시퀀스가 타겟 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈;

플로우 모델을 역으로 수행하여, 음소 길이 예측 모듈에 의해 예측된 길이를 갖는 텍스트 잠재 벡터를 이용하여 참조 음성의 음성 정보를 포함한 음성 정보 벡터를 생성하여 출력하는 플로우 모듈; 및

상기 플로우 모듈로부터 출력된 음성 정보 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;

을 구비하는 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 7

제6항에 있어서, 상기 감정 음성 합성 장치는

상기 임베딩 벡터 검출 모듈에 의해 추출된 운율 임베딩 벡터 및 화자 임베딩 벡터를 상기 플로우 모듈, 음소 길이 예측 모듈 및 디코더 모듈로 제공하는 것을 특징으로 하는 감정 음성 합성 시스템.

청구항 8

(a) 순차적으로 연결된 인코더, 양자화 모듈 및 디코더로 이루어진 오디오 코덱을 이용하여, 오디오 코덱의 양자화 모듈로부터 참조 음성에 대하여 양자화된 벡터를 추출하는 단계; 및

(b) 상기 오디오 코덱의 양자화 모듈로부터 출력된 양자화된 벡터로부터 상기 참조 음성에 대한 운율 정보를 포함하는 운율 임베딩 벡터를 추출하여 출력하는 단계;

를 구비하여, 참조 음성에 대한 양자화된 벡터를 이용하여 운율 임베딩 벡터를 검출하는 것을 특징으로 하는 운율 임베딩 검출 방법.

청구항 9

제8항에 있어서, 상기 (b) 단계는,

(b1) Time-mixing block과 Channel mixing block을 구비하는 Mixer_TTS 구조를 이용하여, 상기 양자화된 벡터에 대하여 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞는 단계; 및

(b2) GRU(Gated Recurrent Unit)를 이용하여, 상기 Mixer_TTS 구조에서 구한 특징값을 문장 단위의 운율 임베

딩 벡터로 변환시켜 출력하는 단계;

를 구비하여, 운율 임베딩 벡터를 생성하여 출력하는 것을 특징으로 하는 운율 임베딩 검출 방법.

청구항 10

텍스트와 정답 음성의 쌍으로 이루어진 데이터 셋을 포함하는 학습 데이터 및 각 화자 정보에 대응되는 화자 임베딩 벡터가 저장된 화자 룩업 테이블을 이용하여 감정 음성 합성 모델을 학습시키는 모델 학습 장치에 있어서,

텍스트에 대응되는 정답 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 정답 음성의 운율 정보를 포함한 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈;

화자 룩업 테이블로부터 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈;

텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈;

상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈;

텍스트에 대응되는 음소 시퀀스가 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈;

음소 길이 예측 모듈을 위한 정답을 만들기 위하여 정답 음성과 텍스트에 대한 정렬 벡터를 추정하는 정렬 검색 모듈;

텍스트에 대한 정답 음성에 대한 Linear Spectrum을 이용하여, 정답 음성의 음성 정보를 포함하는 음성 정답 벡터를 생성하여 출력하는 사후 인코더 모듈;

역연산이 가능한 플로우 모델을 구비하고, 상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터에 대하여 정답 음성의 길이를 갖는 텍스트 정보 벡터를 출력하도록 상기 플로우 모델을 학습시키는 플로우 모듈; 및

상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;

을 구비하여, 운율 임베딩 벡터 및 화자 임베딩 벡터를 감정 음성 합성 모델에 대한 전역 변수로 제공하여 감정 음성 합성 모델을 학습시키는 것을 특징으로 하는 감정 음성 합성 모델 학습 장치.

청구항 11

제10항에 있어서, 상기 모델 학습 장치는,

상기 운율 임베딩 검출 모듈에 의해 추출된 운율 임베딩 벡터 및 상기 화자 임베딩 검출 모듈에 의해 추출된 화자 임베딩 벡터를 상기 플로우 모듈, 정렬 검색 모듈, 음소 길이 예측 모듈, 사후 인코더 모듈 및 디코더 모듈로 제공하는 것을 특징으로 하는 감정 음성 합성 모델 학습 장치.

청구항 12

제10항에 있어서, 상기 운율 임베딩 검출 모듈은,

순차적으로 연결된 인코더, 양자화 모듈 및 디코더로 이루어진 오디오 코덱; 및

상기 오디오 코덱의 양자화 모듈로부터 출력된 양자화된 벡터가 입력되며, 입력된 양자화된 벡터로부터 운율 임베딩 벡터를 추출하여 출력하는 운율 인코더 모듈;을 구비하고,

상기 운율 임베딩 검출 모듈은 텍스트에 대한 정답 음성이 입력되고, 오디오 코덱의 양자화 모듈로부터 상기 입력된 정답 음성에 대한 양자화된 벡터가 출력되고, 상기 양자화된 벡터는 운율 인코더 모듈로 입력되어 운율 임베딩 벡터가 추출되도록 구성된 것을 특징으로 하는 감정 음성 합성 모델 학습 장치.

청구항 13

제12항에 있어서, 상기 운율 인코더 모듈은,

상기 양자화된 벡터에 대하여, 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞는 Time-mixing block과 Channel mixing block을 구비하는 Mixer_TTS 구조; 및

상기 Mixer_TTS 구조에서 구한 특징값을 문장 단위의 운율 임베딩 벡터로 변환시켜 출력하는 GRU(Gated Recurrent Unit);

을 구비하여, 운율 임베딩 벡터를 생성하여 출력하는 것을 특징으로 하는 감정 음성 합성 모델 학습 장치.

청구항 14

입력 텍스트, 참조 음성 및 화자 정보를 입력받는 입력 모듈;

상기 입력된 참조 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 참조 음성의 운율 정보를 포함한 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈;

화자 룩업 테이블로부터 상기 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈;

입력 텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈;

상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈;

입력 텍스트에 대한 음소 시퀀스가 타겟 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈;

플로우 모델을 역으로 수행하여, 음소 길이 예측 모듈에 의해 예측된 길이를 갖는 텍스트 잠재 벡터를 이용하여 참조 음성에 대한 음성 정보를 포함한 음성 정보 벡터를 생성하여 출력하는 플로우 모듈; 및

상기 플로우 모듈로부터 출력된 음성 정보 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;

을 구비하는 것을 특징으로 하는 양자화된 벡터를 이용한 감정 음성 합성 장치.

청구항 15

제14항에 있어서, 상기 감정 음성 합성 장치는

상기 운율 임베딩 벡터 및 화자 임베딩 벡터를 상기 플로우 모듈, 음소 길이 예측 모듈 및 디코더 모듈로 제공하는 것을 특징으로 하는 감정 음성 합성 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 감정 음성 합성 시스템에 관한 것으로서, 더욱 구체적으로는 참조 음성에 대한 양자화된 벡터를 이용하여 운율 임베딩 벡터를 추출하고, 운율 임베딩 벡터 및 화자 임베딩 벡터를 전역 변수로 하여, 입력 텍스트로부터 감정 음성을 합성하는 것을 특징으로 하는 감정 음성 합성 시스템, 감정 음성 합성 장치 및 모델 학습 장치

배 경 기 술

[0002] 음성 합성(Speech Synthesis, Text-to-Speech) 기술은 컴퓨터가 전달하고자 하는 데이터를 인간이 이해할 수 있는 음성 언어로 전달하기 위하여 음성 데이터를 합성해 출력시키는 기술을 말한다. 이러한 음성 합성 기술은 다양한 자연어 처리 기술과 모델들이 도입되고 있다. 일반적으로 은닉 마르코프 모델을 기반으로 한 음성 합성은 음성 주파수, 지속 시간 등이 모두 통계적 예측을 통해 모델링되며, 딥러닝을 기반으로 한 음성 합성은 심층 신경망의 데이터 학습을 바탕으로 음성 데이터가 출력된다.

[0003] 전술한 음성 합성 기술은 장애인들의 편의를 위해 사용되며, 게임, 에니메이션과 같은 엔터테인먼트 산업에서도 활용되고 있으며, 차량 네비게이션이나 ARS와 같은 안내 음성 서비스 등에서도 다양하게 활용되고 있다.

[0004] 이러한 활용이 증대됨에 따라, 자연스러운 음성 합성에 대한 요구도 증대되고 있다. 이러한 요구에 따라, 인공지능은 복합적인 기술들을 사용하여 음성을 학습하는 것과 비슷한 원리로 감정 표현 데이터를 학습하는 기술들이 연구 및 개발되고 있다. 이러한 감정 표현 음성 합성을 위하여, 아래와 같은 다양한 기술들이 제안된 바 있다.

[0005] 한국공개특허공보 제 10-2021-0149608호는 “학습 모델 기반 감정 텍스트-음성 합성 방법 및 장치”에 관한 것으로서, 여러 화자 및 여러 감정이 포함된 음성을 합성하는 단일 학습 모델을 트레이닝하고, 그 학습 모델을 이용하여 감정이 포함된 음성 신호를 생성하는 방법 및 장치에 대한 기술이 개시되어 있다. 전술한 특허 기술은, 입력 텍스트에서 감정 정보를 추출하여 해당 감정에 맞게 감정 음성을 합성하는 기술이다. 해당 특허 기술은 감정 라벨을 예측하여 그 라벨에 따른 감정에 해당하는 운율을 모델링한다. 하지만, 감정을 표현하는 운율은 발화자마다 다르기 때문에, 다양한 표현을 하는 것에 한계가 존재하게 된다.

[0006] 한편, 한국공개특허공보 제 10-2019-0104941호는 “감정 정보 기반의 음성 합성 방법 및 장치”에 관한 것으로서, 대본, 스크립트 등의 상황 설명 정보에 기반하여 감정 정보를 생성하고, 생성된 감정과 텍스트를 결합하여 생동감있는 음성이 출력되도록 하는 음성 합성 장치 및 방법에 대한 기술이 개시되어 있다. 전술한 특허 기술은, 음성으로부터 운율 임베딩 벡터를 학습하는 것이 아니며, 운율을 모델링하기 위해 사전 설정된 여러가지 정보를 입력으로 받고, 사전 설정된 특징값을 이용하여 감정 음성을 모델링하기 때문에, 보다 다양한 감정의 음성을 합성할 수 없다.

[0007] 일반적으로 감정에 따른 운율 정보는 텍스트 뿐만 아니라 음성에 포함된 정보도 매우 중요하다. 하지만, 전술한 특허 기술들에서는 음성에 포함된 운율 정보를 입력받지 않을 뿐만 아니라 음성 정보를 고려하지 않기 때문에, 그 표현력에 한계가 존재하는 문제점이 있다.

[0008] 한편, 추가적인 감정 라벨을 사용하지 않는 비지도 운율 임베딩 벡터 방법을 제안되기도 하였다. 이 기술은, 참조할 감정 음성과 합성할 텍스트를 입력값으로 받아 주어진 음성과 유사한 운율을 표현하는 감정 음성을 합성하도록 학습한다. 이때, 입력으로 주어진 음성을 참조 음성(reference speech)라 하고, mel-spectrogram의 형태로 입력된다. 이 기술에서는 운율 정보를 인코딩하는 과정에서 참조 음성의 내용을 같이 인코딩하게 된다. 운율 임베딩 벡터에 운율 정보 뿐만 아니라 그 내용까지 포함되게 때문에, 그 임베딩을 사용하여 음성합성을 실시하였을 때, 입력 텍스트가 아닌 참조 음성의 언어 정보의 영향을 받게 된다. 그 결과, 합성음의 발음 정확도 등 품질이 저하된다. 이 문제를 content leakage problem이라 한다.

[0009] 이러한 content leakage problem을 해결하기 위하여, 운율 정보를 모델링할 때 mel-spectrogram이 아닌 사전학습 모델의 Latent representation을 사용하는 기술이 제안된 바 있다. Mel-spectrogram에 비해 latent representation은 high-level representation이다. 따라서, style embedding에서 기존의 운율을 복원하는 것이 low-level representation에 비해 어렵기 때문에, 이 기술은 이러한 특징을 이용하여 content leakage problem을 해결하고자 하였다. 하지만, 음성 사전학습 모델에서 추출한 임베딩에는 음성을 생성하는데 필요한 음향 정보가 충분히 포함되어 있지 않다. 따라서, 이 기술은 content leakage problem은 해결되기는 하나, 그 표현력이 저하되는 문제점이 내포되어 있다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 한국공개특허공보 제 10-2021-0149608호

(특허문헌 0002) 한국공개특허공보 제 10-2019-0104941호

비특허문헌

[0011] (비특허문헌 0001) Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis (Wang, Yuxuan, et al., 2018)

(비특허문헌 0002) Fine-grained style control in transformer-based text-to-speech synthesis (Chen, Li-Wei, and Alexander Rudnicky, 2022)

발명의 내용

해결하려는 과제

[0012] 전술한 문제점을 해결하기 위하여, 본 발명은 감정 정보를 나타내기 위하여 참조 음성에 대한 양자화된 벡터로부터 운율 임베딩 벡터를 추출하고, 운율 임베딩 벡터 및 화자 임베딩 벡터를 이용하여 음성 합성하는 것을 특징으로 하는 감정 음성 합성 시스템 및 감정 음성 합성 장치를 제공하는 것을 목적으로 한다.

과제의 해결 수단

[0013] 전술한 기술적 구성을 달성하기 위한 본 발명의 제1 태양에 따른 감정 음성 합성 시스템은, 텍스트와 이에 대한 정답 음성의 쌍으로 이루어진 데이터 셋을 포함하는 학습 데이터 및 각 화자 정보에 대응되는 화자 임베딩 벡터가 저장된 화자 룩업 테이블을 구비하는 데이터 저장 장치; 학습 데이터 및 화자 룩업 테이블을 이용하여 감정 음성 합성 모델을 학습시키는 모델 학습 장치; 및 학습된 감정 음성 합성 모델을 이용하여, 입력 텍스트를 특정 화자의 감정 음성으로 합성하여 출력하는 감정 음성 합성 장치;를 구비하는 감정 음성 합성 시스템에 관한 것으로서, 상기 모델 학습 장치 및 상기 감정 음성 합성 장치는, 참조 음성 또는 정답 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈; 및 화자 룩업 테이블로부터 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈;을 구비하고, 상기 운율 임베딩 벡터 및 상기 화자 임베딩 벡터를 감정 음성 합성 모델에 대한 전역 변수로 제공하는 것을 특징으로 한다.

[0014] 전술한 제1 태양에 따른 감정 음성 합성 시스템에 있어서, 상기 모델 학습 장치는, 텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈; 상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈; 텍스트에 대응되는 음소 시퀀스가 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈; 음소 길이 예측 모듈을 위한 정답을 만들기 위하여 정답 음성과 텍스트에 대한 정렬 벡터를 추정하는 정렬 검색 모듈; 정답 음성에 대한 Linear Spectrum을 이용하여, 정답 음성의 음성 정보를 포함하는 음성 정답 벡터를 생성하여 출력하는 사후 인코더 모듈; 역연산이 가능한 플로우 모델을 구비하고, 상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터에 대하여 정답 음성의 길이를 갖는 텍스트 정보 벡터를 출력하도록 상기 플로우 모델을 학습시키는 플로우 모듈; 및 상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;을 구비한다.

[0015] 전술한 제1 태양에 따른 감정 음성 합성 시스템에 있어서, 상기 모델 학습 장치는, 상기 운율 임베딩 벡터 및 화자 임베딩 벡터를 상기 플로우 모듈, 정렬 검색 모듈, 음소 길이 예측 모듈, 사후 인코더 모듈 및 디코더 모듈로 제공하는 것이 바람직하다.

[0016] 전술한 제1 태양에 따른 감정 음성 합성 시스템에 있어서, 상기 운율 임베딩 검출 모듈은, 순차적으로 연결된 인코더, 양자화 모듈 및 디코더로 이루어진 오디오 코덱; 및 상기 오디오 코덱의 양자화 모듈로부터 출력된 양자화된 벡터가 입력되며, 상기 입력된 양자화된 벡터로부터 운율 정보를 포함하는 운율 임베딩 벡터를 추출하여 출력하는 운율 인코더 모듈;을 구비하고, 상기 운율 임베딩 검출 모듈은 텍스트에 대한 정답 음성이 입력되고, 오디오 코덱의 양자화 모듈로부터 상기 입력된 정답 음성에 대한 양자화된 벡터가 출력되고, 상기 양자화된 벡터는 운율 인코더 모듈로 입력되어 운율 임베딩 벡터가 추출되도록 구성된 것이 바람직하다.

[0017] 전술한 제1 태양에 따른 감정 음성 합성 시스템에 있어서, 상기 운율 인코더 모듈은, 상기 양자화된 벡터에 대하여, 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞는 Time-mixing block과 Channel mixing block을 구비하는 Mixer_TTS 구조; 및 상기 Mixer_TTS 구조에서 구한 특징값을 문장 단위의 운율 임베딩 벡터로 변환시켜 출력하는 GRU(Gated Recurrent Unit);을 구비하여, 운율 임베딩 벡터를 생성하여 출력하는 것이 바람직하다.

[0018] 전술한 제1 태양에 따른 감정 음성 합성 시스템에 있어서, 상기 감정 음성 합성 장치는, 입력 텍스트, 참조 음성 및 화자 정보를 입력받는 입력 모듈; 입력 텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈; 상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈; 입력 텍스트에 대한 음소 시퀀스가 타겟 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈; 플로우 모델을 역으로 수행하여, 음소 길이 예측 모듈에 의해 예측된 길이를 갖는 텍스트 잠재 벡터를 이용하여 음성 정보를 포함한 음성 정보 벡터를 생성하여 출력하는 플로우 모듈; 및 상기 플로우 모듈로부터 출력된 음성 정보 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;을 구비한다.

[0019] 전술한 제1 태양에 따른 감정 음성 합성 시스템에 있어서, 상기 감정 음성 합성 장치는 상기 임베딩 벡터 검출

모듈에 의해 참조 음성으로부터 추출된 운율 임베딩 벡터 및 화자 정보에 따른 화자 임베딩 벡터를 상기 플로우 모듈, 음소 길이 예측 모듈 및 디코더 모듈로 제공하는 것이 바람직하다.

[0020] 본 발명의 제2 태양에 따른 운율 임베딩 검출 방법은, (a) 순차적으로 연결된 인코더, 양자화 모듈 및 디코더로 이루어진 오디오 코덱을 이용하여, 오디오 코덱의 양자화 모듈로부터 참조 음성 또는 정답 음성에 대하여 양자화된 벡터를 추출하는 단계; 및 (b) 상기 오디오 코덱의 양자화 모듈로부터 출력된 양자화된 벡터로부터 상기 참조 음성 또는 정답 음성에 대한 운율 정보를 포함하는 운율 임베딩 벡터를 추출하여 출력하는 단계;를 구비하여, 참조 음성 또는 정답 음성에 대한 양자화된 벡터를 이용하여 운율 임베딩 벡터를 검출한다.

[0021] 전술한 제2 태양에 따른 운율 임베딩 검출 방법에 있어서, 상기 (b) 단계는, (b1) Time-mixing block과 Channel mixing block을 구비하는 Mixer_TTS 구조를 이용하여, 상기 양자화된 벡터에 대하여 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞는 단계; 및 (b2) GRU(Gated Recurrent Unit)를 이용하여, 상기 Mixer_TTS 구조에서 구한 특징값을 문장 단위의 운율 임베딩 벡터로 변환시켜 출력하는 단계;를 구비하여, 운율 임베딩 벡터를 생성하여 출력하는 것이 바람직하다.

[0022] 본 발명의 제3 태양에 따른 감정 음성 합성 모델 학습 장치는, 텍스트와 이에 대한 정답 음성의 쌍으로 이루어진 데이터 셋을 포함하는 학습 데이터 및 각 화자 정보에 대응되는 화자 임베딩 벡터가 저장된 화자 룩업 테이블을 이용하여 감정 음성 합성 모델을 학습시키는 모델 학습 장치에 관한 것으로서, 텍스트에 대한 정답 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 정답 음성의 운율 정보를 포함한 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈; 화자 룩업 테이블로부터 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈; 텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈; 상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈; 텍스트에 대응되는 음소 시퀀스가 정답 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈; 음소 길이 예측 모듈을 위한 정답을 만들기 위하여 정답 음성과 텍스트에 대한 정렬 벡터를 추정하는 정렬 검색 모듈; 정답 음성에 대한 Linear Spectrum을 이용하여, 정답 음성의 음성 정보를 포함하는 음성 정답 벡터를 생성하여 출력하는 사후 인코더 모듈; 역연산이 가능한 플로우 모델을 구비하고, 상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터에 대하여 정답 음성의 길이를 갖는 텍스트 벡터를 출력하도록 상기 플로우 모델을 학습시키는 플로우 모듈; 및 상기 사후 인코더 모듈로부터 출력된 음성 정답 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;을 구비하여, 운율 임베딩 벡터 및 화자 임베딩 벡터를 감정 음성 합성 모델에 대한 전역 변수로 제공하여 감정 음성 합성 모델을 학습시키는 것을 특징으로 한다.

[0023] 전술한 제3 태양에 따른 감정 음성 합성 모델 학습 장치는, 상기 운율 임베딩 검출 모듈에 의해 추출된 운율 임베딩 벡터 및 상기 화자 임베딩 검출 모듈에 의해 추출된 화자 임베딩 벡터를 상기 플로우 모듈, 정렬 검색 모듈, 음소 길이 예측 모듈, 사후 인코더 모듈 및 디코더 모듈로 제공하는 것이 바람직하다.

[0024] 전술한 제3 태양에 따른 감정 음성 합성 모델 학습 장치에 있어서, 상기 운율 임베딩 검출 모듈은, 순차적으로 연결된 인코더, 양자화 모듈 및 디코더로 이루어진 오디오 코덱; 및 상기 오디오 코덱의 양자화 모듈로부터 출력된 양자화된 벡터가 입력되며, 입력된 양자화된 벡터로부터 운율 임베딩 벡터를 추출하여 출력하는 운율 인코더 모듈;을 구비하고, 상기 운율 임베딩 검출 모듈은 텍스트에 대한 정답 음성이 입력되고, 오디오 코덱의 양자화 모듈로부터 상기 입력된 정답 음성에 대한 양자화된 벡터가 출력되고, 상기 양자화된 벡터는 운율 인코더 모듈로 입력되어 운율 임베딩 벡터가 추출되도록 구성된 것이 바람직하다.

[0025] 전술한 제3 태양에 따른 감정 음성 합성 모델 학습 장치에 있어서, 상기 운율 인코더 모듈은, 상기 양자화된 벡터에 대하여, 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞는 Time-mixing block과 Channel mixing block을 구비하는 Mixer_TTS 구조; 및 상기 Mixer_TTS 구조에서 구한 특징값을 문장 단위의 운율 임베딩 벡터로 변환시켜 출력하는 GRU(Gated Recurrent Unit);을 구비하여, 운율 임베딩 벡터를 생성하여 출력하는 것이 바람직하다.

[0026] 본 발명의 제4 태양에 따른 양자화된 벡터를 이용한 감정 음성 합성 장치는, 입력 텍스트, 참조 음성 및 화자 정보를 입력받는 입력 모듈; 상기 입력된 참조 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터를 이용하여 참조 음성의 운율 정보를 포함한 운율 임베딩 벡터를 추출하는 운율 임베딩 검출 모듈; 화자 룩업 테이블로부터 상기 화자 정보에 대응되는 화자 임베딩 벡터를 추출하는 화자 임베딩 검출 모듈; 입력 텍스트에 대한 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터를 추출하여 출력하는 텍스트 인코더 모듈; 상기 텍스트 인코더 모듈로부터 출력된 텍스트 잠재 벡터를 투영하여 출력하는 프로젝션 모듈; 입력

텍스트에 대한 음소 시퀀스가 타겟 음성에서 발음되는 길이를 예측하여 제공하는 음소 길이 예측 모듈; 플로우 모델을 역으로 수행하여, 음소 길이 예측 모듈에 의해 예측된 길이를 갖는 텍스트 잠재 벡터를 이용하여 음성 정보를 포함한 음성 정보 벡터를 생성하여 출력하는 플로우 모듈; 및 상기 플로우 모듈로부터 출력된 음성 정보 벡터를 이용하여 음성 waveform을 합성하여 출력하는 디코더 모듈;을 구비한다.

[0027] 전술한 제4 태양에 따른 양자화된 벡터를 이용한 감정 음성 합성 장치는, 상기 운율 임베딩 벡터 및 화자 임베딩 벡터를 상기 플로우 모듈, 음소 길이 예측 모듈 및 디코더 모듈로 제공하는 것이 바람직하다.

발명의 효과

[0028] 본 발명에 따른 감정 음성 합성 시스템은, 기존의 감정 음성 합성 시스템의 운율 인코더가 Mel-spectrogram 혹은 linear spectrogram을 사용한 것과는 달리, 양자화된 벡터를 운율 인코더의 입력값으로 사용하며, 특히, Neural audio codec에서 multi-stage residual vector quantizer로부터 추출된 vector를 사용하는 것을 특징으로 한다. 이러한 특징에 의하여, 본 발명에 따른 감정 음성 합성 시스템은, 비지도 감정 음성 합성에서의 문제점인 content leakage problem을 해결할 수 있게 된다.

[0029] 본 발명에 따른 감정 음성 합성 시스템에 있어서, Quantized vector를 운율 인코더의 입력으로 사용하는 것은, 음성 사전 학습 모델에서 추출한 운율 임베딩 벡터에 비해 풍부한 음성적 정보를 가지는 효과를 갖는다. Audio codec 모델의 일부분으로 학습된 벡터 양자화 모듈은 복원되었을 때 그 품질을 유지하면서도 low dimension을 가지는 이산 정보로 변환하도록 학습된다. 연속적인 음성 신호를 입력으로 받아 최대한 음성 정보를 유지하면서도 이산적인 정보로 바꾸도록 훈련된다. 이로 인해 양자화된 벡터는 음성 생성에 필요한 음성적 정보를 풍부하게 가지고 있다. 음성 사전 학습 모델은 화자 정보 등 coarse detail만 가지고 있지만, audio codec에서 추출한 양자화된 벡터는 각 음절의 운율 등 fine-detail까지 포함할 수 있기 때문에 더 세밀한 표현이 가능하다.

[0030] 또한, 본 발명에 따른 감정 음성 합성 시스템은 감정 음성 합성 모델을 운율 정보와 화자 정보를 분리하여 학습함으로써, 생성음의 화자 정보가 유실되는 것을 방지한다.

도면의 간단한 설명

[0031] 도 1은 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템을 전체적으로 도시한 구성도이다.

도 2는 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 모델 학습 장치(20)를 개략적으로 도시한 구성도이다.

도 3은 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 운율 임베딩 검출 모듈(212)의 오디오 코덱(213)을 도시한 구성도이다.

도 4는 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 운율 임베딩 검출 모듈(212)의 운율 인코더(214)를 도시한 구성도이다.

도 5는 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 감정 음성 합성 장치(30)를 개략적으로 도시한 구성도이다.

발명을 실시하기 위한 구체적인 내용

[0032] 이하,첨부된 도면을 참조하여 본 발명에 따른 감정 음성 합성 시스템에 대하여 구체적으로 설명한다.

[0033] 도 1은 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템을 개략적으로 도시한 구성도이다. 도 1을 참조하면, 본 발명에 따른 감정 음성 합성 시스템(1)은, 학습 데이터 및 감정 음성 합성 모델을 저장하는 데이터 저장장치(10), 학습 데이터를 이용하여 감정 음성 합성 모델을 학습하는 모델 학습 장치(20), 및 학습이 완료된 감정 음성 합성 모델을 이용하여 입력 텍스트에 대하여 음성을 합성하는 감정 음성 합성 장치(30)를 구비한다.

[0034] 데이터 저장 장치(10)는 학습 데이터 및 화자 룩업 테이블(Speaker Lookup Table)을 구비된다. 상기 학습 데이터는 다양한 발화 스타일을 갖는 다화자 음성 데이터 셋으로 이루어지며, 각 음성 데이터 셋은 텍스트 문장과 이에 대한 정답 음성의 쌍으로 이루어지며, 각 정답 음성은 메타데이터로 화자의 고유 식별 정보(ID)를 포함한다. 상기 화자 룩업 테이블은 화자의 고유 식별 정보(Speaker ID)에 대응되는 화자 정보를 포함하는 화자 임베딩 벡터를 포함한다.

[0035] 다음, 본 발명의 감정 음성 합성 시스템의 모델 학습 장치(20)에 대하여 구체적으로 설명한다.

[0036] 도 2는 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 모델 학습 장치(20)를 개략적으로 도시한 구성도이다. 도 2를 참조하면, 본 발명에 따른 모델 학습 장치(20)는, 데이터 저장 장치에 저장된 학습 데이터 및 화자 룩업 테이블을 이용하여 종단형 감정 음성 합성 모델을 학습시키는 것을 특징으로 한다.

[0037] 상기 모델 학습 장치(20)는, 운율 임베딩 벡터 및 화자 임베딩 벡터를 제공하는 임베딩 벡터 검출 모듈(21), 텍스트 인코더 모듈(22), 프로젝션 모듈(23), 음소 길이 예측 모듈(24), 정렬 검색 모듈(25), 플로우 모듈(26), 사후 인코더 모듈(27) 및 디코더 모듈(28)을 구비한다. 본 발명에 따른 모델 학습 장치는 양자화된 벡터로부터 추출된 운율 임베딩 벡터 및 화자 임베딩 벡터를 전역 조건으로 하여, 사후 인코더 모듈, 디코더 모듈, 플로우 모듈 및 음소 길이 예측 모듈로 제공하는 것을 특징으로 한다.

[0038] 상기 임베딩 벡터 검출 모듈(21)은 화자 임베딩 검출 모듈(210) 및 운율 임베딩 검출 모듈(212)을 구비한다. 상기 화자 임베딩 검출 모듈(210)은 화자 ID에 대응되는 화자 임베딩 벡터를 화자 룩업 테이블로부터 추출하여 출력한다. 상기 임베딩 벡터 검출 모듈(21)에 의해 추출된 운율 임베딩 벡터 및 화자 임베딩 벡터는 음소 길이 예측 모듈(24), 정렬 검색 모듈(25), 플로우 모듈(26), 사후 인코더 모듈(27) 및 디코더 모듈(28)로 제공된다.

[0039] 상기 운율 임베딩 검출 모듈(212)은 오디오 코덱(213) 및 운율 인코더 모듈(214)을 구비하며, 텍스트에 대한 정답 음성이 입력되며, 정답 음성으로부터 양자화된 벡터를 검출하고, 양자화된 벡터를 이용하여 운율 임베딩 벡터를 생성하여 제공한다. 이하, 운율 임베딩 검출 모듈에 대하여 보다 구체적으로 설명한다.

[0040] 도 3은 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 운율 임베딩 검출 모듈(212)의 오디오 코덱(213)을 도시한 구성도이다. 도 3을 참조하면, 상기 오디오 코덱(Audio Codec; 213)은 인코더, multi-stage quantizer로 구성된 양자화 모듈 및 디코더로 이루어진다. 도 3에 예시적으로 도시된 오디오 코덱에 있어서, 인코더는 10개의 CNN Blocks으로 이루어지고, 양자화 모듈은 multi-stage quantizer로 구성되며, 디코더는 10개의 CNN Blocks으로 이루어진다. 오디오 코덱은 음성 waveform이 입력되며. 복원된 음성 waveform이 출력된다. 상기 오디오 코덱은 음성 전송시의 정보 압축을 위해, 인코더와 양자화 모듈을 통해 48khz의 음성을 8 bit rate로 압축하고, 디코더를 통해 손실없이 복원하는 것을 종단형으로 학습한다. 본 발명은 오디오 코덱(213)의 양자화 모듈에서 출력되는 양자화된 벡터(Quantized Vector)를 운율 인코더(214)의 입력값으로 사용하는 것을 특징으로 한다.

[0041] 도 4는 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 운율 임베딩 검출 모듈(212)의 운율 인코더(214)를 도시한 구성도이다. 도 4를 참조하면, 본 발명에 따른 운율 임베딩 검출 모듈(212)의 운율 인코더(214)는 참조 음성 또는 정답 음성에 대한 양자화된 벡터(Quantized Vector)가 입력되며, 상기 참조 음성 또는 정답 음성에 대한 양자화된 벡터로부터 운율 정보를 포함한 운율 임베딩 벡터(Style Embedding Vector)를 생성하여 출력한다.

[0042] 상기 운율 인코더(214)는 Mixer 구조(218) 및 GRU(Gated Recurrent Unit; 219)로 이루어진다. Mixer 구조는 MLP(Multi-Layer Perception)으로 완전 연결 레이어(Fully-Connected Layer)로 이루어지며, Time-mixing block과 Channel mixing block을 구비한다. 상기 Time-mixing block과 Channel mixing block은 음성의 시간축 및 채널축으로 각 특징값의 순서를 랜덤하게 섞어서 정보의 교환이 일어나게 하며, 이를 통해 전체적인 음성 정보를 학습할 수 있게 한다.

[0043] 상기 GRU는 RNN의 한 종류로서, 기본 RNN에 update gate와 reset gate를 추가하여 과거와 현재의 정보를 얼마나 결과값에 반영할지를 결정하는 모듈이다. Mixer 구조에서 encoding한 운율 임베딩 벡터는 음성 프레임마다 고유한 값을 갖게 된다. 그리고, GRU는 Mixer 구조에서 구한 특징값을 총체적으로 나타낼 수 있는 문장 단위의 운율 임베딩 벡터로 변환시키게 된다. 따라서, 전술한 구성을 갖는 운율 인코더는 입력 음성에 대한 양자화된 벡터로부터 해당 음성의 운율 정보를 포함한 운율 임베딩 벡터(Style Embedding Vector)를 생성하여 출력한다.

[0044] 상기 텍스트 인코더 모듈(22)은 텍스트를 이해하는 모듈로서, 텍스트로부터 추출된 음소 시퀀스가 입력된다. 상기 텍스트 인코더 모듈은, 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터(Latent Vector)를 추출하여 출력한다. 텍스트는 G2P 모듈을 통해 음소 시퀀스로 변경되며, 텍스트에 대한 음소 시퀀스가 상기 텍스트 인코더 모듈로 입력된다.

[0045] 상기 프로젝션 모듈(23)은 텍스트 인코더 모듈(22)에서 추출된 텍스트 잠재 벡터가 입력되며, 텍스트와 음성의 관계성을 추출하기 위하여, 텍스트의 벡터보다 더 큰 공간의 잠재 벡터로 투영하여 출력한다.

[0046] 상기 음소 길이 예측 모듈(24)은 텍스트에 대응되는 음소들이 음성에서 발음되는 길이를 예측하는 모듈이다. 상

기 음소 길이 예측 모듈(24)은 텍스트 인코더 모듈(22)에서 출력된 텍스트 잠재 벡터 및 임베딩 벡터 검출 모듈(21)에 의해 제공된 운율 임베딩 벡터 및 화자 임베딩 벡터가 함께 입력되며, 입력 텍스트에 대응되는 음소에 대한 음성의 예측 길이를 출력한다.

[0047] 상기 정렬 검색 모듈(Monotonic Alignment Search Module; 25)은 음소 길이 예측 모듈을 위한 정답을 만들기 위하여 텍스트와 정답 음성에 대한 정렬 벡터(Alignment Vector)를 추정하는 모듈이다. 상기 정렬 검색 모듈(25)은 텍스트의 길이를 갖는 텍스트 잠재 벡터 및 플로우 모듈로부터 출력된 정답 음성의 길이를 갖는 텍스트 정보 벡터가 입력된다. 상기 정렬 검색 모듈은 텍스트에 대한 각 음소가 얼마나 발음되는지에 대한 정답값을 만드는 모듈이므로, 감정 음성 합성 모델을 훈련시키는 모델 학습 장치에서만 사용된다.

[0048] 상기 플로우 모듈(26)은 플로우 모델을 구비한다. 모델 학습 장치의 플로우 모듈은 훈련 과정에서는 텍스트에 대한 정답 음성으로 만든 음성 정답 벡터에 대하여 역연산이 가능한 변형을 거쳐, 정답 음성의 길이를 갖는 텍스트 정보 벡터를 출력하도록, 플로우 모델을 학습한다. 텍스트에 대한 정답 음성에서 추출된 음성 정답 벡터에 대하여 역연산이 가능한 변형을 함으로써, 텍스트의 특징값을 더 갖는 텍스트 정보 벡터들로 변환된다.

[0049] 한편, 상기 플로우 모듈은, 추론 과정에서는 음소 길이 예측 모듈로부터 확장된 텍스트 잠재 벡터를 입력받아서, 훈련에서 학습한 연산을 역으로 수행하고, 해당 역연산을 통해 길이를 제외한 음성의 특징값, 예를 들면 음고, 에너지 등을 벡터에 채워 넣게 넣음으로써, 음성 정보 벡터를 생성한다.

[0050] 상기 사후 인코더 모듈(27)은 음성을 이해하며 음성에 대한 정답 벡터를 만드는 모듈이다. 사후 인코더 모듈은 텍스트에 대한 정답 음성에 대한 Linear Spectrogram이 입력되며, 입력된 Linear Spectrogram을 이용하여 학습시에 사용될 음성 정보를 포함한 음성 정답 벡터를 생성하여 출력한다. 상기 음성 정답 벡터는 텍스트에는 없는 음성만의 고유 정보, 예컨대 음고, 에너지, 음색 등을 포함한 음성 정보 벡터이다.

[0051] 상기 디코더 모듈(28)은 적대적 생성 신경망(Generative Adversarial Network; 'GAN') 구조로 이루어지며, 입력된 음성 정보 벡터로부터 합성된 음성 waveform이 출력된다. 상기 디코더 모듈(28)은, 훈련시에는 사후 인코더 모듈의 마지막 단계의 음성 정답 벡터가 입력되며, 추론시에는 플로우 모듈에서 출력된 음성 정보 벡터가 입력된다.

[0052] 한편, 본 발명에 따른 모델 학습 장치는 훈련된 음성 학습 모델을 추론하는 추론 장치를 더 구비하며, 상기 추론 장치는 후술되는 감정 음성 합성 장치와 동일하게 구성될 수 있다.

[0053] 이하, 본 발명에 따른 감정 음성 합성 시스템의 감정 음성 합성 장치(30)에 대하여 구체적으로 설명한다. 도 5는 본 발명의 바람직한 실시예에 따른 감정 음성 합성 시스템에 있어서, 감정 음성 합성 장치(30)를 개략적으로 도시한 구성도이다. 도 5를 참조하면, 본 발명에 따른 감정 음성 합성 장치(30)는 입력 모듈(도시되지 않음), 임베딩 벡터 검출 모듈(31), 텍스트 인코더 모듈(32), 프로젝션 모듈(33), 음소 길이 예측 모듈(34), 플로우 모듈(36), 디코더 모듈(38)을 구비한다.

[0054] 상기 입력 모듈은 사용자로부터 합성하고자 하는 텍스트 등을 입력받을 수 있는 입력 장치로 구성된다. 상기 입력 모듈을 통해 입력되는 상기 감정 음성 합성 장치의 입력 정보는, 음성 합성하고자 하는 입력 텍스트, 운율 정보에 필요한 참조 음성, 합성음의 화자 정보이다.

[0055] 상기 임베딩 벡터 검출 모듈(31)은, 모델 학습 장치의 임베딩 벡터 검출 모듈과 동일하게 구성되어 화자 임베딩 검출 모듈 및 운율 임베딩 검출 모듈을 구비한다. 상기 운율 임베딩 검출 모듈은 상기 입력된 참조 음성으로부터 양자화된 벡터를 추출하고, 양자화된 벡터로부터 운율 임베딩 벡터를 추출하여 제공한다. 상기 화자 임베딩 검출 모듈은 입력된 화자 정보에 따라 화자 룩업 테이블로부터 화자 임베딩 벡터를 검출하여 제공한다. 상기 임베딩 벡터 검출 모듈은 모델 학습 장치의 임베딩 벡터 검출 모듈과 동일하게 구성되므로, 중복되는 설명은 생략한다. 상기 임베딩 벡터 검출 모듈(31)에 의해 추출된 운율 임베딩 벡터 및 화자 임베딩 벡터는 음소 길이 예측 모듈(34), 플로우 모듈(36) 및 디코더 모듈(38)로 제공된다.

[0056] 상기 텍스트 인코더 모듈(32)은 주어진 텍스트를 이해하는 모듈로서, 입력 텍스트로부터 추출된 음소 시퀀스가 입력된다. 상기 텍스트 인코더 모듈(32)은, 음소 시퀀스로부터 각 음소간의 텍스트적 관계성을 나타내는 텍스트 잠재 벡터(Latent Vector)를 추출하여 출력한다. 상기 텍스트 인코더 모듈(32)은 모델 학습 장치의 텍스트 인코더 모듈(22)과 동일하게 구성되므로, 중복되는 설명은 생략한다.

[0057] 상기 프로젝션 모듈(33)은 상기 텍스트 인코더 모듈(32)에서 추출된 잠재 벡터가 입력되며, 입력 텍스트와 참조 음성의 관계성을 추출하기 위하여, 텍스트의 벡터보다 더 큰 공간의 잠재 벡터로 투영하여 출력한다. 상기 프로

젝션 모듈(33)은 모델 학습 장치의 프로젝션 모듈(23)과 동일하게 구성되므로, 중복되는 설명은 생략한다.

[0058] 상기 음소 길이 예측 모듈(34)은 입력 텍스트에 대응되는 음소 시퀀스가 음성에서 발음되는 길이를 예측하는 모듈이다. 상기 음소 길이 예측 모듈(34)은 텍스트 인코더 모듈(32)에서 출력된 텍스트 잠재 벡터 및 임베딩 벡터 검출 모듈(31)에 의해 제공된 운율 임베딩 벡터 및 화자 임베딩 벡터가 함께 입력되며, 입력 텍스트에 대응되는 음소 시퀀스에 대한 음성의 예측 길이를 출력한다. 상기 음소 길이 예측 모듈(34)은 모델 학습 장치의 음소 길이 예측 모듈(24)과 동일하게 구성되므로, 중복되는 설명은 생략한다.

[0059] 상기 플로우 모듈(36)은 모델 학습 장치에 의해 학습된 플로우 모델을 구비하며, 상기 플로우 모델을 역으로 진행하여 텍스트 잠재 벡터를 음성 정보 벡터로 변환하여 출력한다. 상기 플로우 모듈(36)은 프로젝션 모듈로부터 출력된 텍스트 잠재 벡터가 음소 길이 예측 모듈(34)에 의해 예측된 길이로 확장되어 입력되며, 임베딩 벡터 검출 모듈(31)에 의해 제공된 운율 임베딩 벡터 및 화자 임베딩 벡터도 함께 입력된다. 따라서, 상기 플로우 모듈은 타겟 음성의 길이로 확장된 텍스트 잠재 벡터를 음성 정보 벡터로 변환시켜 출력한다.

[0060] 상기 디코더 모듈(38)은 GAN 구조로 이루어지며, 입력된 음성 정보 벡터로부터 합성된 음성 waveform이 출력된다. 상기 디코더 모듈(38)은, 플로우 모듈(36)에서 출력된 음성 정보 벡터가 입력된다. 상기 디코더 모듈(38)은 모델 학습 장치의 디코더 모듈(28)과 동일하게 구성되므로, 중복되는 설명은 생략한다.

[0061] 이하, 전술한 구성을 갖는 본 발명에 따른 양자화된 벡터를 사용한 감정 음성 합성 모델과 종래의 Spectrum feature를 사용한 감정 감정 음성 합성 모델의 성능을 비교하여 설명한다.

[0062] 성능 평가를 위하여 사용된 평가 데이터는 영어 다화자 음성합성 코퍼스인 VCTK의 음성이며, 훈련에서는 제외한 음성들을 사용하였다. 운율 인코더에서 사용할 참조 음성(reference audio)는 랜덤으로 선정하여 주어졌다. 참조 음성도 훈련에 사용되지 않은 음성들로 구성하였다. 정성 평가에서 평가자들은 모두 영어를 모국어로 하는 화자들로 구성하였다.

[0063] 정성 평가는 Similarity Mean Opinion Score(SMOS)를 사용하였다. SMOS는 평가자들이 1~5점 척도로 운율 정보로 주어진 참조 음성과 합성음의 유사성을 평가하였다. 운율 인코더의 성능을 평가하기 위하여 합성음과 참조 음성의 운율적 유사성을 평가한다. 1점은 정말 유사하지 않음, 5점은 아주 유사함을 나타낸다.

[0064] 정량 평가는 참조 음성과 합성음 각각의 양자화된 벡터의 코사인 유사도를 사용하였다. 코사인 유사도는 아래의 수학식으로 표현된다.

수학식 1

$$\cos(\theta_{VQ}) = \frac{VQ_{ref} \cdot VQ_{\hat{y}}}{\|VQ_{ref}\| \|VQ_{\hat{y}}\|}$$

[0065]

[0066] 코사인 유사도는 1에 가까울 수록 유사하다. 표 1은 운율 인코더에 대한 정성/정량 평가 결과이다. 표 1에서 평가자들은 3가지 기술을 비교한다. GST-FastSpeech2는 종래기술인 GST를 운율 인코더를 사용하였다. Spec-Mixer는 본 발명과 같은 운율 인코더 구조를 사용하지만 그 입력값으로 linear-spectrogram을 사용하는 기술을 지칭한다. 또한, VQ-GST는 종래기술인 GST 와 같은 운율 인코더 구조에 본 발명에 제안하는 양자화된 벡터를 입력값으로 사용하는 기술이다. 이렇게 여러가지 기술 비교를 통해 입력값을 통한 성능 향상의 정도와, 운율 인코더 구조 변경을 통한 성능 향상을 분리하여 분석하였다.

표 1

[0067]

기술 종류	정성평가(SMOS)	정량평가(코싸인 유사도)
GST-FastSpeech2	3.13 ± 0.05	0.61
Spec-Mixer	3.25 ± 0.06	0.69
VQ-GST	3.86 ± 0.07	0.70
본 발명에서 제안한 모듈	3.98 ± 0.04	0.71

[0068] 표 1에서 살펴 보면, 기존 spectrogram based feature를 운율 인코더의 입력으로 사용한 GST-FastSpeech2, Spec-Mixer에 비해 양자화된 벡터를 사용한 VQ-GST와 본 발명의 정성평가에서의 성능이 높게 나타났다. 이는 입력값을 양자화된 벡터를 사용하는 것이 운율 인코더 성능을 높일 수 있다는 것을 보여준다. 정량 평가는 기존 GST-FastSpeech2에 비해 나머지 3기술이 모두 좋은 성능을 보였다. Spec-Mixer는 128차원의 linear spectrogram으로 고차원의 정보를 주었지만, 양자화된 벡터는 8차원으로 상대적으로 저차원의 정보다. 양자화된 벡터를 사용하면 저차원의 정보로도 유사하거나 더 좋은 성능을 낼 수 있음을 알 수 있었다.

[0069] 한편, 입력값에 의한 성능 차이 보다 구조에 의한 성능은 그 차이가 미미했다. 하지만 기존 구조인 CNN에 비해 본 발명에서 사용한 MixerTTS 구조는 그 구조가 간단하다. MixerTTS 구조의 파라미터 개수가 더 적은 것을 고려하면 성능 차이가 미미하다고만은 볼 수 없다.

[0070] 표 2는 각 기술의 파라미터 개수를 비교한 것이다.

표 2

[0071]

기술 종류	모델 파라미터 개수
Spec-Mixer	99,800,000개
VQ-GST	87,100,000개
본 발명에서 제안한 모듈	86,600,000개

[0072] 표 2를 통해, 종래의 기술에 따른 Linear-spectrogram(128차원)을 사용하는 Spec-Mixer에 비해, 본 발명에 따른 양자화된 벡터를 사용했을 때 파라미터 개수가 현저히 줄어드는 것을 확인할 수 있다. 또한, 종래의 기술에 따른 CNN구조를 본 발명에 따른 MixerTTS구조로 변경했을 때 추가적으로 파라미터 개수가 줄어드는 것을 확인할 수 있다. 한편, 종단형 감정 음성 합성 모델의 성능은 합성된 음성의 자연성과 주어진 화자에 대한 표현력 2가지를 평가한다. 감정 음성 합성 모델은 감정 표현 뿐만 아니라 자연스러운 음성을 표현할 수 있어야 한다. 따라서, 본 발명에 따른 감정 음성 합성 시스템에 대하여, 음성의 자연성에 대한 정성 평가도 진행하였다. 또한, 합성된 음성이 운율 정보로 주어진 Reference audio 가 아닌 target speaker의 화자 정보에 맞게 합성되는지 평가하기 위해 화자 표현력을 평가한다. 표 3은 종단형 모듈의 음성의 자연성과 화자 표현력 평가에 대한 결과이다.

[0073] 음성의 자연성에 대한 정성 평가는 Mean Opinion Score(MOS)를 사용하였다. MOS는 평가자들이 1~5점 척도로 음성의 자연성을 평가한다. 1점은 아주 자연스럽지 않음, 5점은 아주 자연스러움을 나타낸다. 화자 표현력에 대한 정량 평가는 화자 임베딩 벡터의 코싸인 유사도를 사용하였다. 수학식 2는 코싸인 유사도를 나타낸다.

수학식 2

$$cos(\theta_{Spk}) = \frac{\mathbf{Spk}_{ref} \cdot \mathbf{Spk}_{\hat{y}}}{\|\mathbf{Spk}_{ref}\| \, \|\mathbf{Spk}_{\hat{y}}\|}$$

[0074]

[0075] 화자 인식 모델에서 화자 임베딩 벡터를 추출하고 이를 원음과 비교하여 원음에 대한 유사도를 평가한다.

표 3

[0076]

기술종류	정성평가(MOS)	정량평가 (화자 코싸인 유사도)
원음	4.53 ± 0.03	-
GST-FastSpeech2	3.68 ± 0.05	0.88
본 발명의 제안 모델	4.00 ± 0.04	0.97

[0077] 비교 기술에 비해 본 발명에서 제안한 모듈이 음성의 자연성, 화자 표현력 모두 높은 성능을 보였다. 특히, 종래 기술에 비해 화자 표현력이 크게 향상되었다. 기존 GST-FastSpeech2에서는 따로 화자 정보를 분리하여 학습하지 않았기 때문이다. 음성의 자연성 또한 종래기술에서는 원음에 비해 그 자연성이 많이 떨어지지만, 본 발명

에서 제안한 모델은 원음에 가까운 성능을 보였다. 기존 GST-FastSpeech2는 단계별로 훈련하였고, 본 발명에서 제안한 모델은 종단형으로 학습하였다. 종단형으로 훈련하여 그 학습의 난이도는 더 높았지만, 더 좋은 성능을 보였다.

[0078] 이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0079] 1 : 감정 음성 합성 시스템

10 : 데이터 저장장치

20 : 모델 학습 장치

30 : 감정 음성 합성 장치

210, 310 : 화자 임베딩 벡터 검출 모듈

212, 312 : 운율 임베딩 벡터 검출 모듈

22, 32 : 텍스트 인코더 모듈

23, 33 : 프로젝션 모듈

24, 34 : 음소 길이 예측 모듈

25 : 정렬 검색 모듈

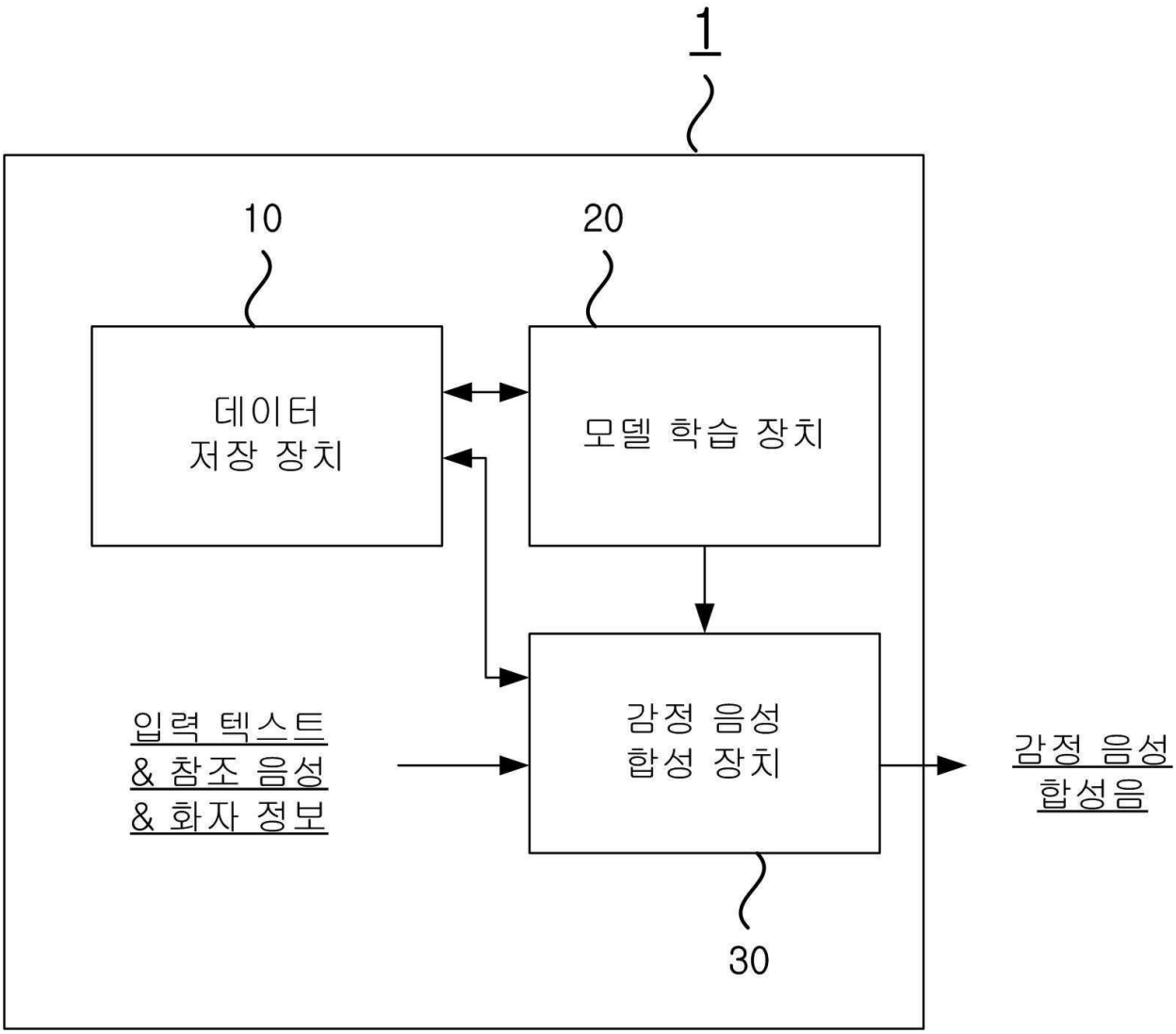
26, 36 : 플로우 모듈

27 : 사후 인코더 모듈

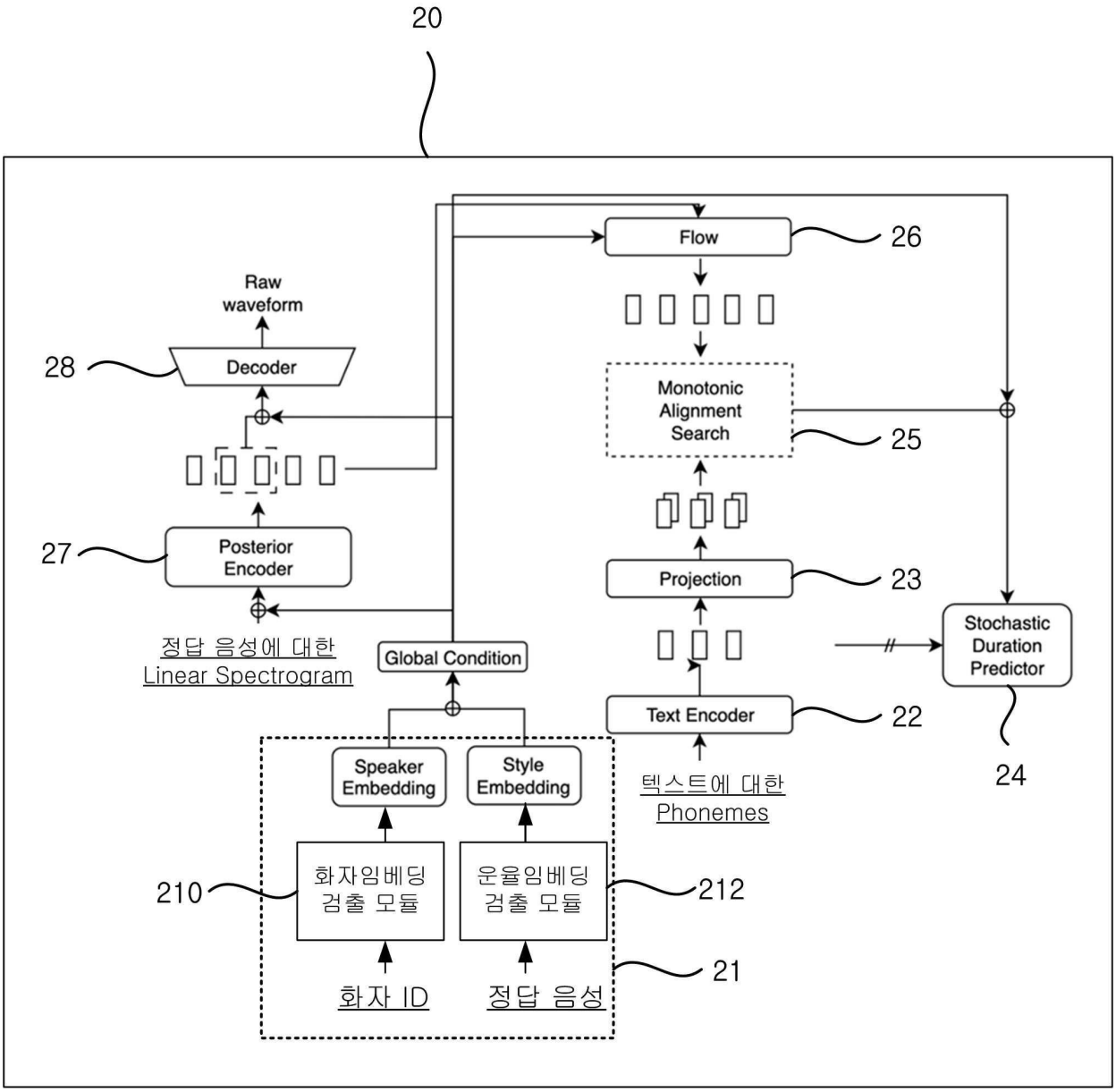
28, 38 : 디코더 모듈

도면

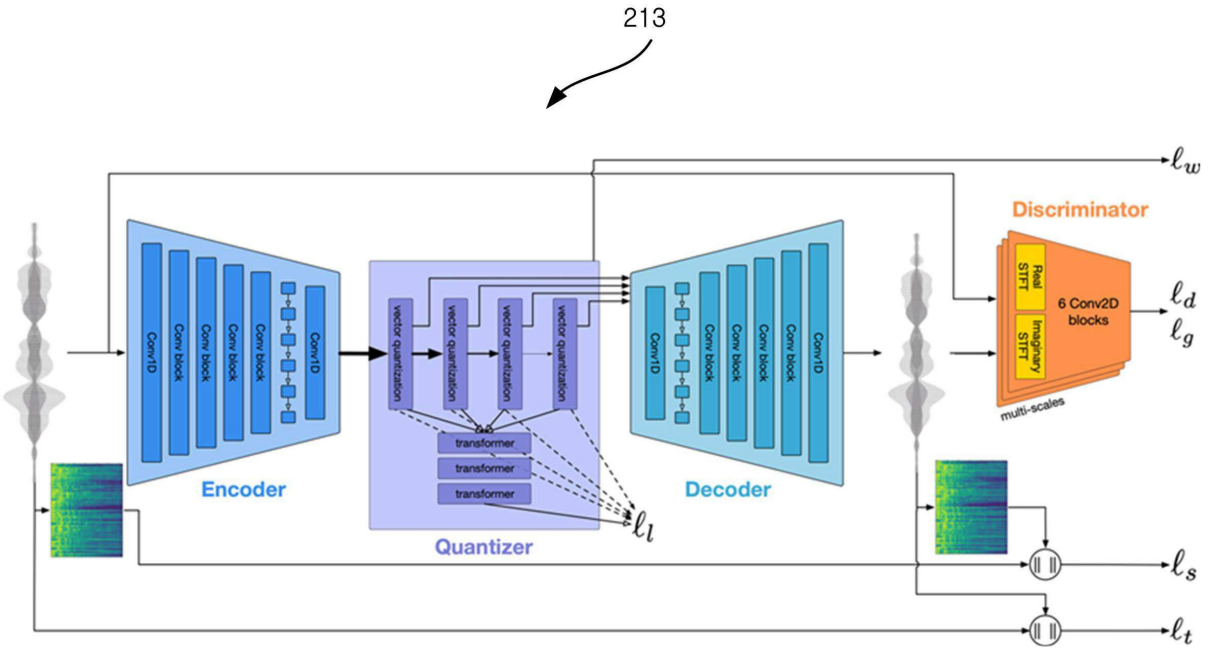
도면1



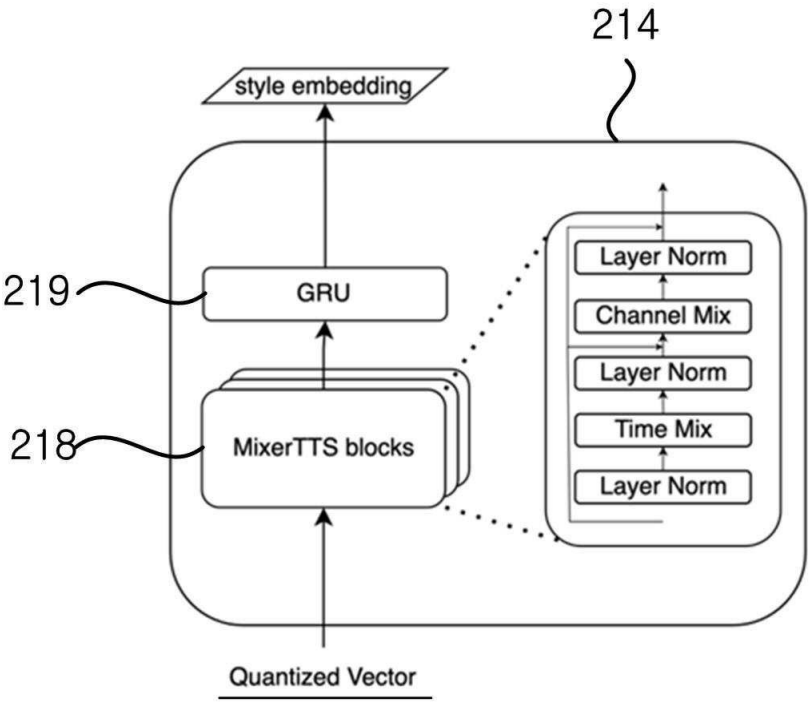
도면2


20
Flow
26
Raw
waveform
Decoder
28
Monotonic
Alignment
Search
25
Posterior
Encoder
27
Projection
23
Stochastic
Duration
Predictor
정답 음성에 대한
Linear Spectrogram
Global Condition
Text Encoder
22
24
Speaker
Embedding
Style
Embedding
텍스트에 대한
Phonemes
210
화자임베딩
검출 모듈
운율임베딩
검출 모듈
212
21
화자 ID
정답 음성

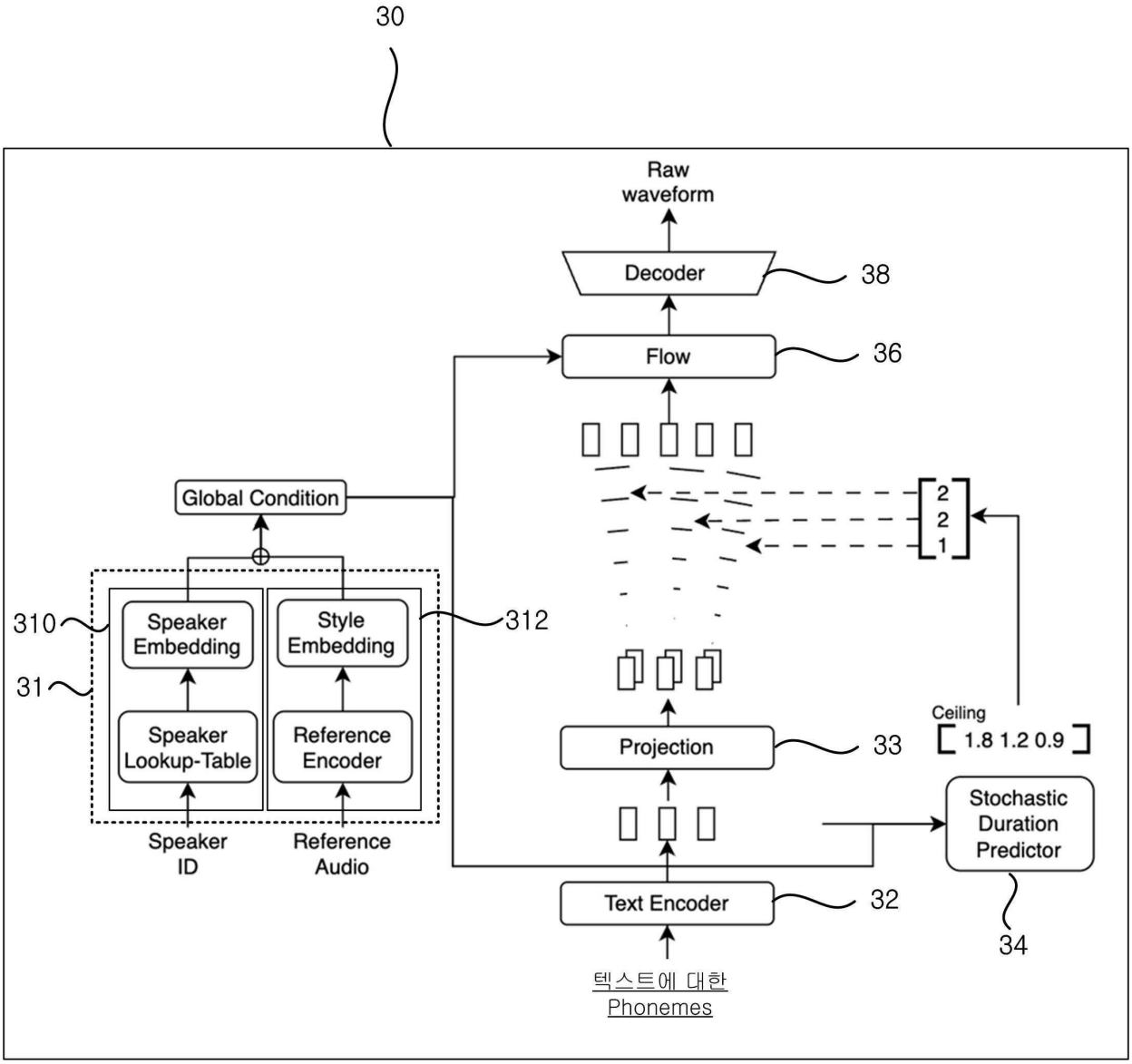
도면3


213
Encoder
Quantizer
Decoder
Discriminator
Real STFT
Imaginary STFT
6 Conv2D blocks
multi-scales
transformer
transformer
transformer

도면4


214
style embedding
219
GRU
218
MixerTTS blocks
Layer Norm
Channel Mix
Layer Norm
Time Mix
Layer Norm
Quantized Vector

도면5


30
Raw waveform
Decoder
38
Flow
36
Global Condition
2
2
1
310
Speaker Embedding
Style Embedding
312
31
Speaker Lookup-Table
Reference Encoder
Speaker ID
Reference Audio
Ceiling
1.8 1.2 0.9
Projection
33
Stochastic Duration Predictor
34
Text Encoder
32
텍스트에 대한 Phonemes