(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년02월21일
(11) 등록번호 10-2639322
(24) 등록일자 2024년02월16일

(51) 국제특허분류(Int. Cl.)
G10L 13/10 (2013.01) *G10L 13/033* (2013.01)
G10L 13/04 (2006.01) *G10L 25/18* (2013.01)
G10L 25/30 (2013.01)
(52) CPC특허분류
G10L 13/10 (2013.01)
G10L 13/033 (2013.01)
(21) 출원번호 10-2022-0079679
(22) 출원일자 2022년06월29일
심사청구일자 2022년06월29일
(65) 공개번호 10-2023-0075340
(43) 공개일자 2023년05월31일
(30) 우선권주장
1020210161696 2021년11월22일 대한민국(KR)
(56) 선행기술조사문헌
KR1020190085882 A*
KR1020210124103 A*
Ohsung Kwon et al., 'Emotional Speech Synthesis Based on Style Embedded Tacotron2 Framework', 34th ITC-CSCC, June 2019.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
이근배
경상북도 포항시 남구 청암로 77
전예진
경상북도 포항시 남구 청암로 77
(74) 대리인
특허법인이상

전체 청구항 수 : 총 18 항 심사관 : 정성윤

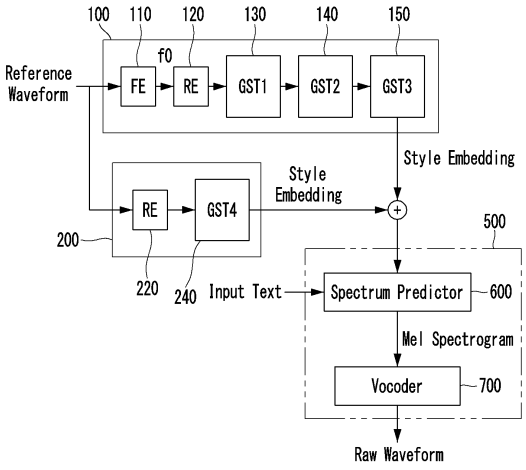
(54) 발명의 명칭 **실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템 및 방법**

(57) 요 약

전역 스타일 토큰을 활용한 딥러닝 기반 종단 간 음성합성 기술을 이용하여 다양한 운율 및 화자 음색 스타일대로 음성을 빠르고 정확하게 합성할 수 있는, 실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템 및 방법이 개시된다. 음성합성 방법은, 입력되는 참조 오디오에서 기본주파수를 추출하여 레퍼런스 인코더의 입력으로

(뒷면에 계속)

대 표 도 - 도1



전달하는 단계, 레퍼런스 인코더에 의해 기본주파수를 인코딩하여 운율 임베딩을 생성하는 단계, 운율 임베딩으로부터 제1 스타일 임베딩을 생성하는 단계, 참조 오디오를 푸리에 변환에 의해 참조 멜 스펙트로그램으로 변환하는 단계, 참조 멜 스펙트로그램을 인코딩하여 스피커 임베딩을 생성하는 단계, 스피커 임베딩으로부터 제2 스타일 임베딩을 생성하는 단계, 제1 스타일 임베딩과 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 음성합성(TTS) 모델의 인코더의 출력과 함께 TTS 모델의 어텐션에 입력하는 단계, 및 TTS 모델에 의해 입력 텍스트에 대하여 음색과 운율이 조합된 음성합성의 오디오를 생성하는 단계를 포함한다.

(52) CPC특허분류
G10L 13/04 (2013.01)
G10L 25/18 (2013.01)
G10L 25/30 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711125943
과제번호 2019-0-01906-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능대학원지원(포항공과대학교)
기 여 율 45/100
과제수행기관명 포항공과대학교 산학협력단
연구기간 2021.01.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1375027305
과제번호 R2021040136-0001
부처명 문화체육관광부
과제관리(전문)기관명 한국콘텐츠진흥원
연구사업명 문화기술연구개발(R&D)
연구과제명 인공지능 및 증가상현실 기반 콘텐츠 메타버스 구축을 통한 R&D 전문인력 양성
기 여 율 45/100
과제수행기관명 한국예술종합학교 산학협력단
연구기간 2021.06.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711126317
과제번호 2020-0-01789-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 High Performance Knowledge System 개발 및 인력양성
기 여 율 10/100
과제수행기관명 동국대학교산학협력단
연구기간 2021.01.01 ~ 2021.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

실시간 음색 및 운율 스타일 복제 가능한 음성합성 방법으로서,

입력되는 참조 오디오에서 기본주파수를 추출하여 레퍼런스 인코더의 입력으로 전달하는 단계;

상기 레퍼런스 인코더에 의해, 상기 기본주파수를 인코딩하여 운율 임베딩을 생성하는 단계;

상기 운율 임베딩으로부터 제1 스타일 임베딩을 생성하는 단계;

상기 참조 오디오를 푸리에 변환에 의해 참조 멜 스펙트로그램으로 변환하는 단계;

상기 참조 멜 스펙트로그램을 인코딩하여 스피커 임베딩을 생성하는 단계;

상기 스피커 임베딩으로부터 제2 스타일 임베딩을 생성하는 단계;

상기 제1 스타일 임베딩과 상기 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 음성합성(text to speech, TTS) 모델의 어텐션에 입력하는 단계; 및

상기 TTS 모델에 의해, 입력되는 텍스트에 대하여 상기 통합 스타일 임베딩에 의한 톤과 운율이 합성된 오디오를 생성하는 단계를 포함하고,

상기 제1 스타일 임베딩을 생성하는 단계는, 제1 전역 스타일 토큰(global style token, GST) 레이어 내지 제3 전역 스타일 토큰 레이어로 구성된 3개 레이어의 계층형(hierarchical) GST를 이용하여 상기 참조 오디오의 운율에 대한 상기 제1 스타일 임베딩을 생성하는, 음성합성 방법.

청구항 2

청구항 1에 있어서,

상기 기본주파수를 추출하는 단계는, 상기 참조 오디오를 일정한 구간의 슬라이딩 윈도우 단위로 자르고, 정규화된 상호 상관함수를 사용하여 상기 참조 오디오의 주파수 분해를 통해 피치(pitch) 윤곽을 계산하고, 여기서 피치는 음의 높낮이에 대응하는, 음성합성 방법.

청구항 3

삭제

청구항 4

청구항 1에 있어서,

상기 제1 내지 제3 GST 레이어들 각각은, 입력되는 임베딩을 각 GST 레이어의 복수의 토큰들로 각각 전달하는 다중 헤드 어텐션과 상기 다중 헤드 어텐션에 연결되는 복수의 토큰들을 구비하고, 상기 입력되는 임베딩과 각 토큰 간의 유사도를 측정하고 각각의 토큰들의 가중 합으로 스타일 임베딩을 생성하며,

상기 제1 내지 제3 GST 레이어들은, 현재 GST 레이어의 상기 복수의 토큰들과 이전 GST 레이어의 상기 복수의 토큰들을 연결하는 잔여 커넥션(residual connection)을 구비하는, 음성합성 방법.

청구항 5

청구항 1에 있어서,

상기 제2 스타일 임베딩을 생성하는 단계는, 제4 GST 레이어를 이용하여 상기 참조 오디오의 음색에 대한 상기 제2 스타일 임베딩을 생성하는, 음성합성 방법.

청구항 6

청구항 1에 있어서,

상기 오디오를 생성하는 단계는,

상기 TTS 모델의 인코더에 의해, 상기 인코더에 입력되는 텍스트로부터 특징 정보를 추출하는 단계;

상기 어텐션에 의해, 상기 특징 정보를 매 시점마다 상기 TTS 모델의 디코더에서 사용할 어텐션 얼라인(alignment) 정보로 매핑하고 매핑된 어텐션 얼라인 정보와 상기 통합 스타일 임베딩을 상기 디코더로 전달하는 단계; 및

상기 디코더에 의해, 상기 어텐션의 어텐션 정보와 이전 시점의 멜 스펙트로그램을 이용하여 상기 통합 스타일 임베딩이 합성된 현재 시점의 멜 스펙트로그램을 생성하는 단계를 포함하는 음성합성 방법.

청구항 7

청구항 6에 있어서,

상기 오디오를 생성하는 단계는, MelGAN(Mel generative adversarial networks)에 의해, 상기 현재 시점의 멜 스펙트로그램으로부터 오디오를 생성하는 단계를 더 포함하는, 음성합성 방법.

청구항 8

청구항 1에 있어서,

상기 참조 멜 스펙트로그램에 대응하는 타겟 멜 스펙트로그램과 상기 TTS 모델에서 생성된, 상기 입력되는 텍스트와 상기 참조 오디오에 대한 예측 멜 스펙트로그램을 비교하여 상기 타겟 멜 스펙트로그램과 상기 예측 멜 스펙트로그램과의 차이 또는 상기 차이에 대응하는 손실(loss)을 구하는 단계; 및

상기 차이 또는 손실이 미리 설정된 수준 또는 기준값 이하가 될 때까지 상기 TTS 모델을 훈련시키는 단계를 더 포함하는, 음성합성 방법.

청구항 9

실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템으로서,

입력되는 참조 오디오에서 기본주파수를 추출하는 추출부;

상기 기본주파수를 입력받고 상기 기본주파수를 인코딩하여 운율 임베딩을 생성하는 제1 레퍼런스 인코더;

상기 운율 임베딩으로부터 상기 참조 오디오의 운율에 대한 제1 스타일 임베딩을 생성하는 계층형 전역 스타일 토큰 레이어들;

상기 참조 오디오를 푸리에 변환에 의해 참조 멜 스펙트로그램으로 변환하는 변환부;

상기 참조 멜 스펙트로그램을 인코딩하여 스피커 임베딩을 생성하는 제2 레퍼런스 인코더;

상기 스피커 임베딩으로부터 상기 참조 오디오의 음색에 대한 제2 스타일 임베딩을 생성하는 단일 전역 스타일 토큰 레이어; 및

상기 제1 스타일 임베딩과 상기 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 어텐션으로 입력받고 입력되는 텍스트에 대하여 상기 통합 스타일 임베딩에 의한 톤과 운율이 합성된 오디오를 생성하는 음성합성(text to speech, TTS) 모델을 포함하고,

상기 계층형 전역 스타일 토큰 레이어들은, 제1 전역 스타일 토큰(global style token, GST) 레이어 내지 제3 전역 스타일 토큰 레이어로 구성된 3개 레이어의 계층형 GST로 구성되는, 음성합성 시스템.

청구항 10

청구항 9에 있어서,

상기 추출부는, 상기 참조 오디오를 일정한 구간의 슬라이딩 윈도우 단위로 자르고, 정규화된 상호 상관함수를 사용하여 유효 프레임을 구분하고, 상기 유효 프레임 내 상기 참조 오디오의 주파수 분해를 통해 피치(pitch) 윤곽을 계산하며, 여기서 피치는 상기 참조 오디오의 음의 높낮이에 대응하는, 음성합성 시스템.

청구항 11

삭제

청구항 12

청구항 9에 있어서,

상기 계층형 전역 스타일 토큰 레이어들의 각 GST 레이어는 다중 헤드 어텐션과 상기 다중 헤드 어텐션에 연결되는 복수의 토큰들을 구비하고,

상기 제1 GST 레이어는 상기 운율 임베딩과 각 토큰 간의 유사도를 측정하고 상기 복수의 토큰들의 가중 합으로 스타일 임베딩을 생성하고,

상기 제1 GST 레이어에서 생성된 스타일 임베딩은 상기 제2 GST 레이어 및 제3 GST 레이어를 순차적으로 통과하여 제1 스타일 임베딩으로서 출력되는, 음성합성 시스템.

청구항 13

청구항 9에 있어서,

상기 계층형 전역 스타일 토큰 레이어들은, 상기 제1 GST 레이어와 상기 제2 GST 레이어와의 제1 쌍과 상기 제2 GST 레이어와 상기 제3 GST 레이어와의 제2 쌍은 현재 레이어의 토큰들이 이전 레이어의 토큰들과 연결되는 잔여 커넥션(residual connection)을 구비하는, 음성합성 시스템.

청구항 14

청구항 9에 있어서,

상기 참조 멜 스펙트로그램에 대응하는 타겟 멜 스펙트로그램과, 상기 TTS 모델에서 생성된, 상기 입력되는 텍스트와 상기 참조 오디오에 대한 예측 멜 스펙트로그램을 비교하여 상기 타겟 멜 스펙트로그램과 상기 예측 멜 스펙트로그램과의 차이 또는 상기 차이에 대응하는 손실(loss)을 구하는 학습관리부를 더 포함하며,

상기 학습관리부는 상기 차이 또는 손실이 미리 설정된 수준 또는 기준값 이하가 될 때까지 상기 TTS 모델을 훈련시키는, 음성합성 시스템.

청구항 15

청구항 9에 있어서,

상기 TTS 모델은, 스펙트로그램 예측기 및 보코더를 구비하고,

상기 스펙트로그램 예측기는 상기 입력되는 텍스트와 상기 통합 스타일 임베딩을 토대로 멜 스펙트로그램을 생성하고,

상기 보코더는 상기 멜 스펙트로그램으로부터 합성 파형에 대응하는 파형 샘플을 생성하는, 음성합성 시스템.

청구항 16

청구항 15에 있어서,

상기 스펙트로그램 예측기는 인코더, 어텐션 및 디코더를 포함하고,

상기 인코더는 상기 입력되는 텍스트로부터 특징 정보를 추출하고,

상기 어텐션은, 상기 특징 정보를 매 시점마다 상기 디코더에서 사용할 어텐션 얼라인 정보(attention alignment information)로 매핑하고 매핑된 어텐션 얼라인 정보와 상기 통합 스타일 임베딩을 상기 디코더로 전달하며,

상기 디코더는, 상기 어텐션의 어텐션 정보와 이전 시점의 멜 스펙트로그램을 이용하여 상기 통합 스타일 임베딩이 합성된 현재 시점의 멜 스펙트로그램을 생성하는, 음성합성 시스템.

청구항 17

청구항 16에 있어서,

상기 스펙트로그램 예측기는 상기 인코더의 입력단에 위치하는 전처리부를 더 포함하며, 상기 전처리부는 입력되는 텍스트를 음절 단위로 분리하고, 분리된 음절을 원핫 인코딩(one-hot encoding)을 통해 정수로 표현하는, 음성합성 시스템.

청구항 18

청구항 17에 있어서,

상기 인코더는, 문자 임베딩(character embedding) 생성 유닛, 상기 문자 임베딩 생성 유닛에 연결되는 3 컨볼루션 레이어들(3 Conv layers), 및 상기 3 컨볼루션 레이어들에 연결되는 양방향 LSTM(bidirectional long short term memory)을 구비하고,

상기 문자 임베딩 생성 유닛은 상기 전처리부로부터 받은 정수 시퀀스를 매트릭스 형태로 변환하고,

상기 3 컨볼루션 레이어들은 매트릭스 형태의 정보를 축약하고,

상기 양방향 LSTM은 축약된 매트릭스 형태의 정보를 인코더 특징 정보로 변환하며, 여기서 상기 인코더 특징 정보는 하나의 고정된 크기로 압축된 컨텍스트 벡터를 포함하는, 음성합성 시스템.

청구항 19

청구항 16에 있어서,

상기 어텐션은 상기 TTS 모델의 보코더를 통해 출력될 음성 발음이 상기 입력되는 텍스트의 순차적인 순서대로 진행되도록 상기 디코더의 타임-스텝에 따라 상기 어텐션 얼라인 정보를 상기 인코더 특징 정보에 추가하는, 음성합성 시스템.

청구항 20

청구항 15에 있어서,

상기 보코더는, 멜겐(MelGAN: Mel generative adversarial networks)인, 음성합성 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 다양한 운율 및 화자 음색 스타일대로 음성을 빠르고 정확하게 합성하는 음성합성 시스템에 관한 것으로, 전역 스타일 토큰(Global Style Token, GST)을 활용한 딥러닝 기반 종단 간(end-to-end) 음성합성 기술을 이용하는 실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 음성합성(Text to Speech, TTS)의 궁극적인 목표는 입력된 텍스트를 정확하고 자연스러운 목소리로 읽어주는 기술이며, 사람이 직접 발성하는 모든 분야에 적용 가능하다. 음성합성의 대표적인 응용 분야에는 네비게이션, 오디오북, 인공지능 비서 등이 있다.

[0003] 또한, 급격히 발전하는 딥러닝 기술을 활용하여 합성된 음성의 품질은 매우 향상되었으며, 종단 간(end-to-end) 음성합성 모델들이 제안되면서 별도의 지식이나 작업 없이 비교적 간단한 모델 학습이 가능해졌다.

[0004] 이러한 발전에도 불구하고, 현재의 딥러닝 기반의 종단 간 음성합성 시스템은 일상대화에서 쉽게 들을 수 있는 음성의 운율적 정보 예컨대 감정, 억양, 어조 등이나 화자의 음색을 그대로 따라하지 못하는 한계가 있으므로 추가적인 연구가 요구된다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 전술한 종래 기술의 요구에 부응하기 위해 도출된 것으로, 본 발명의 목적은,실시간으로 다양한 운율

적 요소와 화자의 음색을 복제할 수 있는 실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템 및 방법을 제공하는데 있다.

[0006] 본 발명의 다른 목적은, 전역 스타일 토큰(Global Style Token, GST) 기반 모듈을 사용하여 음색과 운율 스타일을 실시간 복제하여 합성할 수 있는 음성합성 시스템 및 방법을 제공하는데 있다.

[0007] 본 발명의 또 다른 목적은, 실시간 음색 및 운율 스타일 복제와 함께 감정 제어 가능한 음성합성 시스템 및 방법을 제공하는데 있다.

과제의 해결 수단

[0008] 상기 기술적 과제를 달성하기 위한 본 발명의 일 측면에 따른 음성합성 방법은, 실시간 음색 및 운율 스타일 복제 가능한 음성합성 방법으로서, 입력되는 참조 오디오(reference waveform)에서 기본주파수를 추출하여 레퍼런스 인코더의 입력으로 전달하는 단계; 상기 레퍼런스 인코더에 의해, 상기 기본주파수를 인코딩하여 운율 임베딩을 생성하는 단계; 상기 운율 임베딩으로부터 제1 스타일 임베딩을 생성하는 단계; 상기 참조 오디오를 푸리에 변환에 의해 참조 멜 스펙트로그램으로 변환하는 단계; 상기 참조 멜 스펙트로그램을 인코딩하여 스피커 임베딩을 생성하는 단계; 상기 스피커 임베딩으로부터 제2 스타일 임베딩을 생성하는 단계; 상기 제1 스타일 임베딩과 상기 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 음성합성(text to speech, TTS) 모델의 어텐션에 입력하는 단계; 및 상기 TTS 모델에 의해, 입력되는 텍스트에 대하여 상기 통합 스타일 임베딩에 의한 톤과 운율이 합성된 오디오를 생성하는 단계를 포함한다.

[0009] 일실시예에서, 상기 기본주파수를 추출하는 단계는, 상기 참조 오디오를 일정한 구간의 슬라이딩 윈도우 단위로 자르고, 정규화된 상호 상관함수를 사용하여 상기 참조 오디오의 주파수 분해를 통해 피치(pitch) 윤곽을 계산하는 프로세스를 포함할 수 있다. 여기서 피치는 상기 참조 오디오의 음의 높낮이에 대응할 수 있다.

[0010] 일실시예에서, 상기 제1 스타일 임베딩을 생성하는 단계는, 제1 전역 스타일 토큰(global style token, GST) 레이어 내지 제3 전역 스타일 토큰 레이어로 구성된 3개 레이어의 계층형(hierarchical) GST를 이용하여 상기 참조 오디오의 운율에 대한 상기 제1 스타일 임베딩을 생성하도록 구성될 수 있다.

[0011] 일실시예에서, 상기 제1 내지 제3 GST 레이어들 각각은, 입력되는 임베딩을 각 GST 레이어의 복수의 토큰들로 각각 전달하는 다중 헤드 어텐션과 상기 다중 헤드 어텐션에 연결되는 복수의 토큰들을 구비하고, 상기 입력되는 임베딩과 각 토큰 간의 유사도를 측정하고 각각의 토큰들의 가중 합으로 스타일 임베딩을 생성하도록 구성될 수 있다.

[0012] 일실시예에서, 상기 제1 내지 제3 GST 레이어들은, 현재 GST 레이어의 상기 복수의 토큰들과 이전 GST 레이어의 상기 복수의 토큰들을 연결하는 잔여 커넥션(residual connection)을 구비할 수 있다.

[0013] 일실시예에서, 상기 제2 스타일 임베딩을 생성하는 단계는, 제4 GST 레이어를 이용하여 상기 참조 오디오의 음색에 대한 상기 제2 스타일 임베딩을 생성하도록 구성될 수 있다.

[0014] 일실시예에서, 상기 오디오를 생성하는 단계는, 상기 TTS 모델의 인코더에 의해, 상기 인코더에 입력되는 텍스트의 문자(characters)로부터 특징 정보를 추출하는 단계; 상기 어텐션에 의해, 상기 특징 정보를 매 시점마다 상기 TTS 모델의 디코더에서 사용할 어텐션 얼라인(alignment) 정보로 매핑하고 매핑된 어텐션 얼라인 정보와 상기 통합 스타일 임베딩을 상기 디코더로 전달하는 단계; 및 상기 디코더에 의해, 상기 어텐션의 어텐션 정보와 이전 시점의 멜 스펙트로그램을 이용하여 상기 통합 스타일 임베딩이 합성된 현재 시점의 멜 스펙트로그램을 생성하는 단계를 포함하도록 구성될 수 있다.

[0015] 일실시예에서, 상기 오디오를 생성하는 단계는, MelGAN(Mel generative adversarial networks)에 의해, 상기 현재 시점의 멜 스펙트로그램으로부터 오디오를 생성하는 단계를 더 포함할 수 있다.

[0016] 일실시예에서, 음성합성 방법은, 상기 참조 멜 스펙트로그램에 대응하는 타겟 멜 스펙트로그램과 상기 TTS 모델에서 생성된, 상기 입력되는 텍스트와 상기 참조 오디오에 대한 예측 멜 스펙트로그램을 비교하여 상기 타겟 멜 스펙트로그램과 상기 예측 멜 스펙트로그램과의 차이 또는 상기 차이에 대응하는 손실(loss)을 구하는 단계; 및 상기 차이 또는 손실이 미리 설정된 수준 또는 기준값 이하가 될 때까지 상기 TTS 모델을 훈련시키는 단계를 더 포함할 수 있다.

[0017] 상기 기술적 과제를 달성하기 위한 본 발명의 다른 측면에 따른 음성합성 시스템은, 실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템으로서, 입력되는 참조 오디오(reference waveform)에서 기본주파수를 추출하는

추출부; 상기 기본주파수를 입력받고 상기 기본주파수를 인코딩하여 운율 임베딩을 생성하는 제1 레퍼런스 인코더; 상기 운율 임베딩으로부터 상기 참조 오디오의 운율에 대한 제1 스타일 임베딩을 생성하는 계층형 전역 스타일 토큰 레이어들; 상기 참조 오디오를 푸리에 변환에 의해 참조 멜 스펙트로그램으로 변환하는 변환부; 상기 참조 멜 스펙트로그램을 인코딩하여 스피커 임베딩을 생성하는 제2 레퍼런스 인코더; 상기 스피커 임베딩으로부터 상기 참조 오디오의 음색에 대한 제2 스타일 임베딩을 생성하는 단일 전역 스타일 토큰 레이어; 및 상기 제1 스타일 임베딩과 상기 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 어텐션으로 입력받고 입력되는 텍스트에 대하여 상기 통합 스타일 임베딩에 의한 톤과 운율이 합성된 오디오를 생성하는 음성합성(text to speech, TTS) 모델을 포함한다.

[0018] 일실시예에서, 상기 추출부는, 상기 참조 오디오를 미리 설정된 일정한 구간의 슬라이딩 윈도우 단위로 자르고, 정규화된 상호상관함수를 사용하여 유효 프레임을 구분하고, 상기 유효 프레임 내 상기 참조 오디오의 주파수 분해를 통해 피치(pitch) 윤곽을 계산하도록 구성될 수 있다. 여기서 피치는 상기 참조 오디오의 음의 높낮이에 대응할 수 있다.

[0019] 일실시예에서, 상기 추출부는 YIN 방법을 이용하여 피치 윤곽에 대응하는 기본 주파수를 추정할 수 있다. 여기서 YIN은 자기상관(autocorrelation)과 소거(cancellation) 사이의 상호작용을 암시하는 동양철학의 음(Yin)과 양(Yang)의 합성어이다.

[0020] 일실시예에서, 상기 계층형 전역 스타일 토큰 레이어들은, 제1 전역 스타일 토큰(global style token, GST) 레이어 내지 제3 전역 스타일 토큰 레이어로 구성된 3개 레이어의 계층형 GST 레이어들로 구성될 수 있다.

[0021] 일실시예에서, 상기 계층형 전역 스타일 토큰 레이어들의 각 GST 레이어는 다중 헤드 어텐션과 상기 다중 헤드 어텐션에 연결되는 복수의 토큰들을 구비할 수 있다. 여기서, 상기 제1 GST 레이어는 상기 운율 임베딩과 각 토큰 간의 유사도를 측정하고 각각의 토큰들의 가중 합으로 스타일 임베딩을 생성하고, 상기 제1 GST 레이어에서 생성된 스타일 임베딩은 상기 제2 GST 레이어 및 제3 GST 레이어를 순차적으로 통과하여 제1 스타일 임베딩으로서 생성될 수 있다.

[0022] 일실시예에서, 상기 계층형 전역 스타일 토큰 레이어들은, 상기 제1 GST 레이어와 상기 제2 GST 레이어와의 제1 쌍과 상기 제2 GST 레이어와 상기 제3 GST 레이어와의 제2 쌍은 현재 레이어의 토큰들이 이전 레이어의 토큰들과 연결되는 잔여 커넥션(residual connection)을 구비할 수 있다.

[0023] 일실시예에서, 음성합성 시스템은, 상기 참조 멜 스펙트로그램에 대응하는 타겟 멜 스펙트로그램과, 상기 TTS 모델에서 생성된, 상기 입력되는 텍스트와 상기 참조 오디오에 대한 예측 멜 스펙트로그램을 비교하여 상기 타겟 멜 스펙트로그램과 상기 예측 멜 스펙트로그램과의 차이 또는 상기 차이에 대응하는 손실(loss)을 구하는 학습관리부를 더 포함할 수 있다. 학습관리부는 상기 차이 또는 손실이 미리 설정된 수준 또는 기준값 이하가 될 때까지 상기 TTS 모델을 훈련시킬 수 있다.

[0024] 일실시예에서, 상기 TTS 모델은, 스펙트로그램 예측기 및 보코더를 구비할 수 있다. 여기서, 상기 스펙트로그램 예측기는 상기 입력되는 텍스트와 상기 통합 스타일 임베딩을 토대로 멜 스펙트로그램을 생성하고, 상기 보코더는 상기 멜 스펙트로그램으로부터 합성 파형에 대응하는 파형 샘플을 생성할 수 있다.

[0025] 일실시예에서, 상기 스펙트로그램 예측기는 인코더, 어텐션 및 디코더를 포함할 수 있다. 상기 인코더는 상기 입력되는 텍스트의 문자(characters)로부터 특징 정보를 추출할 수 있다. 상기 어텐션은, 상기 특징 정보를 매 시점마다 상기 디코더에서 사용할 어텐션 얼라인(alignment) 정보로 매핑하고 매핑된 어텐션 얼라인 정보와 상기 통합 스타일 임베딩을 상기 디코더로 전달할 수 있다. 그리고, 상기 디코더는, 상기 어텐션의 어텐션 정보와 이전 시점의 멜 스펙트로그램을 이용하여 상기 통합 스타일 임베딩이 합성된 현재 시점의 멜 스펙트로그램을 생성할 수 있다.

[0026] 일실시예에서, 상기 스펙트로그램 예측기는 상기 인코더의 입력단에 위치하는 전처리부를 더 포함할 수 있다. 상기 전처리부는 입력되는 텍스트를 음절 단위로 분리하고, 분리된 음절을 원핫 인코딩(one-hot encoding)을 통해 정수로 표현하도록 구성될 수 있다.

[0027] 일실시예에서, 상기 인코더는, 문자 임베딩(character embedding) 유닛, 상기 문자 임베딩 생성 유닛에 연결되는 3개의 컨볼루션 레이어들(3 Conv layers), 및 상기 3개의 컨볼루션 레이어들에 연결되는 양방향 LSTM(bidirectional long short term memory)을 구비할 수 있다. 상기 문자 임베딩 생성 유닛은 상기 전처리부로부터 받은 정수 시퀀스를 매트릭스 형태로 변환할 수 있다. 상기 컨볼루션 레이어들은 매트릭스 형태의 정보를 축약할 수 있다. 그리고 상기 양방향 LSTM은 축약된 매트릭스 형태의 정보를 인코더 특징 정보로 변환할 수

있다. 여기서 상기 인코더 특징 정보는 하나의 고정된 크기로 압축된 컨텍스트 벡터(context vector)를 포함할 수 있다.

[0028] 일실시예에서, 상기 어텐션은 상기 TTS 모델의 보코더를 통해 출력될 음성 발음이 상기 입력되는 텍스트의 순차적인 순서대로 진행되도록 상기 디코더의 타임-스텝에 따라 상기 어텐션 얼라인 정보를 상기 인코더 특징 정보에 추가하도록 구성될 수 있다.

[0029] 일실시예에서, 상기 보코더는, 멜겐(MelGAN: Mel generative adversarial networks)일 수 있다.

[0030] 일실시예에서, 음성합성 시스템은, 상기 참조 오디오를 인코딩하여 감정 임베딩을 생성하는 제3 레퍼런스 인코더; 및 상기 감정 임베딩으로부터 상기 참조 오디오에 포함된 감정 정보에 대한 제3 스타일 임베딩을 생성하는 또 다른 전역 스타일 토큰 레이어를 더 포함할 수 있다. 제3 스타일 임베딩은 상기 제1 스타일 임베딩 및 상기 제2 스타일 임베딩과 함께 통합 스타일 임베딩에 포함되어 상기 TTS 모델의 어텐션에 입력될 수 있다.

[0031] 일실시예에서, 제3 레퍼런스 인코더는 상기 참조 오디오로부터 SFTF를 통해 변환된 멜 스펙트로그램(이하 '참조 멜 스펙트로그램')이 입력되는 3개의 합성곱 신경망(three convolutional neural networks, 3 CNN), 상기 3개의 합성곱 신경망에 연결되는 2개의 잔여 블록들(2 residual blocks) 및 상기 2개의 잔여 블록들에 연결되는 4개 레이어들의 수정 알렉스넷(modified AlexNet)으로 구성될 수 있다.

[0032] 일실시예에서, 상기 제3 레퍼런스 인코더 및 상기 또 다른 전력 스타일 토큰 레이어와 결합되는 TTS 모델은, 참조 오디오의 스타일 트랜스퍼 및 스타일 트랜스퍼의 강도를 조절하도록 구성되고, 리컨스트럭션 로스(reconstruction loss)만으로 학습을 진행할 수 있다.

[0033] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 측면에 따른 음성합성 시스템은, 전술한 실시예들 중 어느 하나의 실시간 음색 및 운율 스타일 복제 가능한 음성합성 방법을 구현하기 위한 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램을 포함할 수 있다.

[0034] 상기 기술적 과제를 달성하기 위한 본 발명의 또 다른 측면에 따른 음성합성 시스템은, 전술한 실시예들 중 어느 하나의 실시간 음색 및 운율 스타일 복제 가능한 음성합성 방법을 구현하기 위한 프로그램을 기록한 컴퓨터 판독 가능한 기록매체를 포함할 수 있다.

발명의 효과

[0035] 전술한 음성합성 시스템 및 방법의 구성에 의하면, 전역 스타일 토큰(global style token, GST) 기반 모듈을 사용하여 참조 오디오의 음색이나 운율 스타일을 복제하여 입력 텍스트에 대한 음성합성 결과에 효과적으로 반영할 수 있다.

[0036] 또한, 본 발명에 의하면, 입력 텍스트에 대하여 전역 스타일 토큰을 기반으로 음색, 운율 및 감정의 다양한 스타일을 반영한 음성합성 파형을 효과적으로 생성할 수 있다.

[0037] 또한, 본 발명에 의하면, 실시간으로 참조 오디오의 다양한 운율적 요소, 화자의 음색, 또는 화자의 감정을 복제하여 입력 텍스트에 대한 음성합성 출력에 실시간으로 추가할 수 있는 음성합성 시스템 및 방법을 제공할 수 있다.

도면의 간단한 설명

[0038] 도 1은 본 발명의 일 실시예에 따른 실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템(이하 간략히 '음성합성 시스템')의 전체 구성에 대한 개략적인 블록도이다.

도 2는 비교예의 음성합성 시스템에 대한 블록도이다.

도 3은 도 1의 음성합성 시스템에 채용할 수 있는 운율 모듈에 대한 구성도이다.

도 4는 도 1의 음성합성 시스템에 채용할 수 있는 전체 구조에 대한 상세 구성도이다.

도 5는 본 발명의 다른 실시예에 따른 음성합성 시스템의 전체 구성도이다.

도 6은 도 5의 음성합성 시스템에 채용할 수 있는 감정 모듈을 설명하기 위한 블록도이다.

도 7은 도 6의 감정 모듈에 채용할 수 있는 레퍼런스 인코더에 대한 상세 구성도이다.

도 8은 본 발명의 또 다른 실시예에 따른 음성합성 시스템에 대한 개략적인 구성도이다.

도 9는 도 8의 음성합성 시스템에 적용될 수 있는 주요 동작을 설명하기 위한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0039] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하여 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.

[0040] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. '및/또는'이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0041] 본 출원의 실시예들에서, 'A 및 B 중에서 적어도 하나'는 'A 또는 B 중에서 적어도 하나' 또는 'A 및 B 중 하나 이상의 조합들 중에서 적어도 하나'를 의미할 수 있다. 또한, 본 출원의 실시예들에서, 'A 및 B 중에서 하나 이상'은 'A 또는 B 중에서 하나 이상' 또는 'A 및 B 중 하나 이상의 조합들 중에서 하나 이상'을 의미할 수 있다.

[0042] 어떤 구성요소가 다른 구성요소에 '연결되어' 있다거나 '접속되어' 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 '직접 연결되어' 있다거나 '직접 접속되어'있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0043] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, '포함한다' 또는 '가진다' 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0044] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0045] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0047] 도 1은 본 발명의 일 실시예에 따른 실시간 음색 및 운율 스타일 복제 가능한 음성합성 시스템(이하 간략히 '음성합성 시스템')의 전체 구성에 대한 개략적인 블록도이다.

[0048] 도 1을 참조하면, 음성합성 시스템은 운율 모듈(100), 음색 모듈(200), 스펙트럼 예측기(spectrum predictor, 600) 및 보코더(vocoder, 700)를 포함하고, 운율 모듈(100)의 제1 스타일 임베딩(style embedding)과 음색 모듈(200)의 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 적용하여 입력 텍스트(input text)에 특정 운율 및 음색이 가미된 최초 음성 파형(raw waveform)을 출력할 수 있다.

[0049] 운율 모듈(100)은 추출부(110), 제1 레퍼런스 인코더(reference encoder, RE, 120), 제1 전역 스타일 토큰 레이어(global style token layer, GST1, 130), 제2 GST 레이어(GST2, 140) 및 제3 GST 레이어(GST3, 150)를 구비한다.

[0050] 추출부(110)는 입력되는 참조 파형(reference waveform)인 참조 오디오에서 기본주파수(fundamental frequency, f0)를 추출한다. 추출부(110)는 참조 오디오에서 운율과 관련된 특징을 추출하는 특징 추출부(feature extractor, FE)로 지칭될 수 있다. 여기서 운율은 참조 오디오의 매 기준 시점의 음높이를 지칭하거나 음높이와 강세의 변화를 지칭할 수 있다.

[0051] 제1 레퍼런스 인코더(120)는 추출부(110)으로부터 받은 기본주파수의 참조 오디오를 인코딩하여 운율 임베딩을 생성한다. 제1 레퍼런스 인코더(120)는 일정한 구간의 슬라이딩 윈도우 단위들을 배치놈(batchNorm)을 가진 3개의 합성곱 신경망을 통해 필터링하고, 2개의 잔여 블록들을 통해 앞서 학습된 정보를 이용하여 가중치들을 계산한 후, 4개의 레이어들로 수정된 알렉스넷으로 가중치를 최적화하도록 이루어질 수 있다. 여기서 운율 임베딩은 참조 오디오의 운율을 설명하는 매개변수로서, 학습이 완료된 후에 생성되는 운율 임베딩은 참조 오디오를 가장 잘 설명하는 최적의 매개변수일 수 있다.

[0052] 제1 GST 레이어(GST1, 130), 제2 GST 레이어(GST2, 140) 및 제3 GST 레이어(GST3, 150)는 계층형(hierarchical) GST 구조를 형성한다. 즉, 제1 내지 제3 GST 레이어들(130, 140, 150)은 기재된 순서대로 임베딩이 순차적으로 처리하도록 나열된 구조를 가진다. 각 GST 레이어는 다중 헤드 어텐션과 상기 다중 헤드 어텐션에 연결되는 복수의 토큰들을 구비한다. 이러한 계층형 GST 레이어들(이하 간략히 '계층형 GST')는 제1 레퍼런스 인코더(120)로부터 입력되는 운율 임베딩을 적절하게 그리고 효과적으로 처리하기 위한 것이다.

[0053] 계층형 GST에서, 제1 GST 레이어(130)는 운율 임베딩과 자신의 각 토큰 간의 유사도를 측정하고 토큰들의 가중합(weithted sum)으로 스타일 임베딩을 생성한다. 즉, 제1 GST 레이어(130)는 운율 임베딩을 k(5 이상의 임의의 자연수)개의 전역 스타일 토큰들(GSTs)의 가중 합으로 변화시켜 표현할 수 있다. 이 과정에서 k개의 GSTs은 이들의 조합이 다양한 운율 스타일을 표현할 수 있도록 학습된다. 특히, 본 실시예에서는, 제1 GST 레이어(130)에서 학습된 스타일 임베딩을 제2 GST 레이어(140) 및 제3 GST 레이어(150)에서 다시 학습시킴으로써, 운율 임베딩의 다양한 표현에 대하여 처리 시간과 시스템 복잡도를 고려하여 최대한 효과적으로 학습할 수 있도록 구성된다. 계층형 GST의 학습 결과는 제1 스타일 임베딩으로서 출력된다.

[0054] 더욱이, 본 실시예에서는, 운율 임베딩의 더욱 효과적인 학습을 위해, 계층형 GST가 제1 GST 레이어(130)와 제2 GST 레이어(140)의 제1 쌍과 제2 GST 레이어(140)와 제3 GST 레이어(150)의 제2 쌍 각각에 대하여 잔여 커넥션(residual connection)을 구비하도록 구성될 수 있다.

[0055] 이러한 잔여 커넥션은, 3개의 GST 레이어들(130, 140, 150)을 사용하여 운율 임베딩의 다양한 표현을 학습할 때, 데이터 처리 흐름 상에서 상대적으로 후단에 배치된 현재 GST 레이어가 상대적으로 전단에 배치된 이전 GST 레이어를 학습 결과를 이용하도록 하기 위한 것으로, 운율 임베딩의 학습 효과를 극대화하는데 기여할 수 있다.

[0056] 한편, 본 실시예에서는 계층형 GST가 3개의 GST 레이어들로 이루어지는 것이 가장 바람직한 형태임을 설명하였지만, 본 발명은 그러한 구성으로 한정되지 않고, 계층형 GST를 2개, 4개 혹은 5개의 GST 레이어들로 구성하는 경우에도, 본 실시예와 유사한 효과를 얻을 수 있음은 물론이다. 5개를 초과하는 GST 레이어들의 구성은 처리 시간과 시스템 복잡도 등을 고려하여 제한될 수 있다.

[0057] 음색 모듈(200)은 제2 레퍼런스 인코더(RE, 220) 및 제4 GST 레이어(GST4, 240)을 구비하고, 제2 레퍼런스 인코더(220)에서 변환된 스피커 임베딩을 제4 GST 레이어(240)에서 학습하고 제2 스타일 임베딩을 출력하도록 구성될 수 있다.

[0058] 제2 레퍼런스 인코더(220)의 입력으로는 참조 오디오를 특정 길이의 슬라이딩 윈도우 단위들로 잘라 푸리에 변환을 통해 만들어진 스펙트로그램(spectrogram)이 사용될 수 있다. 이러한 스펙트로그램을 만드는 과정은 간략히 STFT(short-time fourier transform)로 지칭될 수 있다. 스펙트로그램은 멜 필터 뱅크(Mel-filter bank)를 이용한 로그 변환(log transform)을 통해 멜 스펙트로그램으로 변환될 수 있다.

[0059] 제4 GST 레이어(240)는, 스피커 임베딩을 효과적으로 학습하기 위해 사용된다. 제4 GST 레이어(240)는 스피커 임베딩과 자신의 각 토큰 간의 유사도를 측정하고 토큰들의 가중 합(weithted sum)으로 스타일 임베딩(제2 스타일 임베딩)을 생성한다. 즉, 제4 GST 레이어(240)는 스피커 임베딩을 k(5 이상의 임의의 자연수)개의 전역 스타일 토큰들(GSTs)의 가중 합으로 변화시켜 표현할 수 있으며, 이 과정에서 k개의 GSTs은 이들의 조합이 다양한 음색 스타일을 표현할 수 있도록 학습된다.

[0060] 스펙트럼 예측기(spectrum predictor, 600)와 보코더(vocoder, 700)는 음성합성(text to speech, TTS) 모델(500)을 형성한다. 본 실시예에서 TTS 모델(500)은 제1 스타일 임베딩과 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 받고, 입력되는 텍스트 즉, 입력 텍스트(input text)에 대한 음성합성 결과를 생성할 때 통합 스타일 임베딩에 대응하는 음색과 운율을 반영하도록 구성된다.

[0061] 여기서, 스펙트럼 예측기(600)는 통합 스타일 임베딩을 반영하는 입력 텍스트에 대한 멜 스펙트로그램을 생성한다. 그리고, 보코더(700)는 멜 스펙트로그램으로부터 최초 음성 파형(raw waveform)을 출력한다. 최초 음성 파

형은 합성 파형(synthesized waveform) 또는 파형 샘플(waveform samples)로서 WAV(waveform audio format) 등의 특정 포맷을 가질 수 있으나, 이에 한정되지는 않는다.

[0062] 스펙트럼 예측기(600)는 텍스트가 입력되는 인코더, 인코더에 연결되는 어텐션, 어텐션과 상호 연결되는 디코더를 구비할 수 있다.

[0063] 보코더(700)는 스펙트로그램이나 멜 스펙트로그램으로부터 파형(waveform) 신호 또는 음성 신호를 생성하는 모듈이다. 보코더(700)는 웨이브넷(WaveNet) 보코더, 멜겐(MelGAN: Mel generative adversarial network) 등을 사용할 수 있다. 멜겐(MelGAN)을 사용하면, 웨이브넷 보코더 등 다른 보코더보다 참조 오디오의 음색이나 운율을 잘 표현하는 음성 신호를 생성할 수 있다.

[0065] 도 2는 비교예의 음성합성 시스템에 대한 블록도이다.

[0066] 도 2를 참조하면, 비교예의 음성합성 시스템은, 스펙트럼 예측기(10), 레퍼런스 인코더(20) 및 웨이브넷 보코더((WaveNet vocoder, 90)로 구성된다. 스펙트럼 예측기(10)는 인코더(encoder, 30), 어텐션(attention, 50) 및 디코더(decoder, 70)로 구성된다.

[0067] 앞서 설명한 본 실시예의 음성합성 시스템(도 1 참조)과 대비할 때, 비교예의 음성합성 시스템은, 레퍼런스 인코더(20)에서 참조 오디오(reference audio)로부터 생성된 참조 임베딩과 인코더(30)의 특징 정보를 합하여 어텐션(50)에 입력될 때, 어텐션(50)과 디코더(70)가 이들의 협업을 통해 멜 스펙트로그램을 생성하고, 웨이브넷 보코더(90)가 멜 스펙트로그램으로부터 음성합성 결과인 발과(speech)를 생성하도록 구성된다.

[0068] 전술한 비교예의 음성합성 시스템과 대비할 때, 본 실시예의 음성합성 시스템은, 참조 오디오의 운율을 효과적으로 표현하기 위해, 참조 오디오의 기본주파수를 제1 레퍼런스 인코더의 입력으로 사용하고, 제1 레퍼런스 인코더에서 생성된 운율 임베딩을 계층형 GST의 복수의 GST 레이어들(특히, 제1 내지 제3 GST 레이어들)에 입력하여 운율을 학습하도록 구성됨을 알 수 있다.

[0069] 또한, 본 실시예의 음성합성 시스템은, 참조 오디오의 음색을 효과적으로 표현하기 위해, 참조 오디오의 스펙트로그램을 제2 레퍼런스 인코더의 입력으로 사용하고, 제2 레퍼런스 인코더에서 생성된 스피커 임베딩을 단일 GST 레이어(제4 GST 레이어)에 입력하여 음색을 학습하도록 구성됨을 알 수 있다.

[0070] 또한, 본 실시예의 음성합성 시스템은, 운율 학습 결과로 얻은 제1 스타일 임베딩과 음색 학습 결과로 얻은 제2 스타일 임베딩을 합하여 스펙트럼 예측기의 인코더 출력과 합하거나 스펙트럼 예측기의 어텐션에 인코더의 출력과 함께 입력되도록 구성됨을 알 수 있다.

[0071] 더욱이, 본 실시예의 음성합성 시스템은, 후술되는 바와 같이, 참조 오디오의 감정을 효과적으로 표현하기 위해, 참조 오디오의 스펙트로그램을 제3 레퍼런스 인코더의 입력으로 사용하고, 제3 레퍼런스 인코더에서 생성된 감정 임베딩을 단일 GST 레이어(제5 GST 레이어)에 입력하여 상기의 감정을 학습하도록 구성됨을 알 수 있다. 여기서, 제5 GST 레이어에서 출력되는 제3 스타일 임베딩은 제1 및 제2 스타일 임베딩들과 합쳐져 스펙트럼 예측기를 포함하는 TTS 모델에 전달될 수 있다.

[0072] 또한, 본 실시예의 음성합성 시스템은 보코더로서 멜겐(MelGAN)을 사용함으로써 더욱 효과적으로 음색과 운율의 스타일이나 음색, 운율 및 감정 스타일이 복제된 음성합성 결과를 생성하도록 구성됨을 알 수 있다.

[0074] 도 3은 도 1의 음성합성 시스템에 채용할 수 있는 운율 모듈에 대한 구성도이다.

[0075] 도 3을 참조하면, 운율 모듈(100)은 추출부(FE, 110), 레퍼런스 인코더(reference encoder, 120), 제1 GST 레이어(GST1 또는 GST layer 1, 130) 내지 제N GST 레이어(GST layer N, 160)를 구비한다. 여기서 레퍼런스 인코더(120)은 제1 레퍼런스 인코더에 대응하고, N은 2 내지 5 중 어느 하나의 자연수일 수 있다.

[0076] 운율 모듈(100)은 추출부(110)에 의해 입력되는 참조 파형(reference waveform) 또는 참조 오디오로부터 기본주파수(f0)를 추출하고, 제1 레퍼런스 인코더(120)에 의해 기본주파수로부터 운율 임베딩(prosody embedding)을 생성하고, 제1 내지 제N GST 레이어들(130, 160)의 계층형 GST에 의해 운율 임베딩을 학습하고 제1 스타일 임베딩(style embedding)을 출력하도록 구성된다.

[0077] 제1 GST 레이어(130)는 다중 헤드 어텐션(multi-head attention, 132)과 다중 헤드 어텐션(132)에 연결되는 복수의 토큰들(token 1 내지 token k)(134)을 구비한다. 여기서, k는 5 이상의 임의의 자연수일 수 있다. 제1 GST 레이어(130)는 운율 임베딩과 각 토큰(token 1 내지 token k) 간의 유사도를 측정하고 이 토큰들의 가중 합으로 스타일 임베딩을 생성한다. 제1 GST 레이어(130)에서 생성된 스타일 임베딩은 제2 GST 레이어의 입력으로 전달

될 수 있다.

[0078] 위의 경우와 유사하게, 제N GST 레이어(160)는 다중 헤드 어텐션(162)과 다중 헤드 어텐션(162)에 연결되는 복수의 토큰들(token 1 내지 token k)(164)을 구비한다. 제N GST 레이어(160)는 입력단에 위치하는 다른 GST 레이어에서 받은 스타일 임베딩과 각 토큰(token 1 내지 token k) 간의 유사도를 측정하고 이 토큰들의 가중 합으로 스타일 임베딩을 생성한다. 이 스타일 임베딩은 운율 모듈(100)에 최종적으로 출력되는 제1 스타일 임베딩이 된다.

[0079] 한편, 상기의 N이 2인 경우, 제2 GST 레이어는 제1 GST 레이어(130)로부터 받은 스타일 임베딩과 자신의 복수의 토큰들 각각과의 유사도를 측정하고 이 토큰들의 가중 합으로 스타일 임베딩을 생성할 수 있다. 이때, 제1 GST 레이어(130)의 복수의 토큰들(134)은 제2 GST 레이어의 복수의 토큰들과 잔여 커넥션(residual connection, 170)으로 연결될 수 있고, 그에 의해 제2 GST 레이어는 제1 GST 레이어(130)의 복수의 토큰들(134)의 가중치를 토대로 학습을 병렬적으로 수행할 수 있다.

[0080] 그리고, 상기의 N이 3인 경우, 위의 N이 2인 경우에 더하여, 제3 GST 레이어는 제2 GST 레이어로부터 받은 스타일 임베딩과 자신의 복수의 토큰들 각각과의 유사도를 측정하고 이 토큰들의 가중 합으로 스타일 임베딩을 생성할 수 있다. 이때, 제2 GST 레이어의 복수의 토큰들은 제3 GST 레이어의 복수의 토큰들과 또 다른 잔여 커넥션(residual connection, 170)으로 연결될 수 있고, 그에 의해 제3 GST 레이어는 제2 GST 레이어의 토큰들의 가중치를 토대로 학습을 효과적으로 수행할 수 있다.

[0082] 도 4는 도 1의 음성합성 시스템에 채용할 수 있는 전체 구조에 대한 상세 구성도이다.

[0083] 도 4를 참조하면, 음성합성 시스템은 운율 모듈(100), 음색 모듈(200), 스펙트럼 예측기(spectrum predictor, 600) 및 보코더로서 멜겐(MelGAN, 700a)를 포함하고, 운율 모듈(100)의 제1 스타일 임베딩(style embedding)과 음색 모듈(200)의 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 적용하여 입력 텍스트(input text)에 특정 운율 및 음색이 부가된 파형 샘플(waveform samples)을 출력할 수 있다.

[0084] 운율 모듈(100)은 추출부(FE, 110), 제1 레퍼런스 인코더(reference encoder, 120), 제1 스타일 토큰 레이어(style token layer 1, 130), 제2 스타일 토큰 레이어(style token layer 2, 140) 및 제3 스타일 토큰 레이어(style token layer 3, 150)를 구비한다. 각 스타일 토큰 레이어는 전역 스타일 토큰(GST) 레이어에 대응된다.

[0085] 운율 모듈(100)에서, 추출부(110)는 참조 파형인 참조 오디오로부터 기본주파수(fundamental frequency, f0)를 추출한다. 참조 오디오는 미리 지정된 데이터셋에 있는 오디오일 수 있다. 또한, 참조 오디오는 학습(traning)시 원본, 타겟 또는 참값을 나타내는 그라운드 트루(ground truth, GT)일 수 있고 추정(inference)시 원하는 스타일이 포함된 오디오일 수 있다. 기본주파수는 제1 레퍼런스 인코더(120)의 입력으로 사용된다.

[0086] 제1 레퍼런스 인코더(120)는 기본주파수로부터 운율 임베딩(prosody embedding)을 생성한다. 제1 레퍼런스 인코더(120)는 6개의 2D 합성곱 신경망(convolutional neural networks, CNN)과 1개의 게이트 순환 유닛(gated recurrent unit, GRU)으로 구성될 수 있다. 운율 임베딩은 기본주파수가 제1 레퍼런스 인코더를 통과한 결과로서, 정해진 길이의 임베딩이고, 계층형 GST의 입력으로 사용된다.

[0087] 3개의 스타일 토큰 레이어들(130, 140, 150)로 구성된 계층형 GST는 운율 임베딩을 학습하여 제1 스타일 임베딩을 생성할 수 있다. 각 스타일 토큰 레이어(130, 140, 150)는 다중헤드 어텐션(multi-head attention)을 구비한다. 계층형 GST는, 다중헤드 어텐션을 통해 제1 레퍼런스 인코더(120)에서 출력된 운율 임베딩과 각 스타일 토큰 사이의 유사도가 측정되면, 이 스타일 토큰들의 가중 합으로 스타일 임베딩을 생성한다.

[0088] 운율 모듈(100)의 스타일 임베딩은, 음색 모듈(200)의 스타일 임베딩과 다르게, 첫 번째로 생성된 스타일 임베딩이 두 개의 스타일 토큰 레이어를 추가로 통과할 때, 첫 번째 스타일 토큰 레이어와 두 번째 스타일 토큰 레이어와의 사이 및 두 번째 스타일 토큰 레이어와 세 번째 스타일 토큰 레이어와의 사이 각각에 잔여 커넥션(residual connection)을 추가로 사용하여 참조 오디오의 운율 특징에 대한 학습 효과를 극대화하도록 구성될 수 있다.

[0089] 음색 모듈(200)은 STFT 유닛(210), 제2 레퍼런스 인코더(220) 및 제4 스타일 토큰 레이어(style token layer 4, 240)을 구비한다. STFT 유닛(210)은, 참조 오디오를 별도로 가공한 스펙트로그램이나 멜 스펙트로그램을 제2 레퍼런스 인코더(220)의 입력 데이터로 준비하는 경우에, 생략될 수 있다. 제4 스타일 토큰 레이어(240)는 제4 GST 레이어에 대응된다.

[0090] 음색 모듈(200)에서, STFT 유닛(210)은 참조 오디오를 푸리에 변환에 의해 변환한 멜 스펙트로그램을 제2 레퍼

런스 인코더(220)의 입력(input)으로 제공한다. 제2 레퍼런스 인코더(220)는 기본주파수가 아닌 멜 스펙트로그램을 입력받아 스피커 임베딩을 생성한다. 제4 GST 레이어(240)를 통과한 스피커 임베딩은 제2 스타일 임베딩으로 출력된다.

[0091] 제4 스타일 토큰 레이어(240)는 스피커 임베딩을 학습하여 제2 스타일 임베딩을 생성할 수 있다. 제4 스타일 토큰 레이어(240)는 다중헤드 어텐션(multi-head attention, 242) 및 복수의 토큰들(244)을 구비한다. 복수의 토큰들(244)은 제1 토큰(token 1) 내지 제5 토큰(token 5)을 포함할 수 있고, 각 토큰은 스타일 토큰으로 지칭될 수 있다.

[0092] 제4 스타일 토큰 레이어(240)는, 다중헤드 어텐션(242)을 통해 제2 레퍼런스 인코더(220)에서 출력된 스피커 임베딩과 각 스타일 토큰 사이의 유사도가 측정되면, 이 스타일 토큰들(244)의 가중 합으로 스타일 임베딩(제2 스타일 임베딩)을 생성할 수 있다.

[0093] 음율 모듈(100)에서 총 3개의 스타일 토큰 레이어들을 통과한 스타일 임베딩(제1 스타일 임베딩)과 음색 모듈(200)의 스타일 임베딩(제2 스타일 임베딩)은 합쳐지고(concatenate) 음성합성 시스템의 스펙트럼 예측기(600)의 인코더(encoder)의 출력과 결합된다.

[0094] 본 실시예의 음성합성 시스템은 좁은 의미에서 스펙트럼 예측기(600)와 보코더로서 멜겐(MelGAN, 700a)으로 구성될 수 있고, 넓은 의미에서 음율 모듈(100)과 음색모듈(200)을 더 포함할 수 있다. 여기서, 스펙트럼 예측기(600)는 인코더(610), 어텐션(620) 및 디코더(630)를 구비한다.

[0095] 즉, 스펙트럼 예측기(600)는 입력 텍스트(input text)를 받는 인코더(610), 인코더(610)의 출력단에 연결되는 어텐션(620), 어텐션(620)과 상호 연결되는 디코더(630)를 구비한다. 이러한 스펙트럼 예측기(600)는 타코트론2(Tacotron2)로 구현될 수 있다.

[0096] 전술한 경우, 음성합성 시스템은 seq2seq 음성합성 시스템으로서 두 단계를 통해 텍스트를 음성으로 합성할 수 있다. 첫 번째 단계에서 입력 텍스트로부터 멜 스펙트로그램을 생성하고, 두 번째 단계에서 보코더(vocoder)를 사용하여 멜 스펙트로그램으로부터 음성을 합성한다.

[0097] 본 실시예의 스펙트럼 예측기(600)는, 일반적인 타코트론2(Tacotron2)와 비교할 때, 제1 및 제2 스타일 임베딩들을 합한 통합 스타일 임베딩이 인코더(610)의 출력과 결합되어 어텐션(620)에 입력되는 점에서 차이가 있다.

[0098] 좀더 구체적으로, 스펙트럼 예측기(600)에 있어서, 인코더(610)는 입력 텍스트를 인코딩하여 512 차원의 특징(이하 '특징 정보')으로 변환한다. 인코더(610)는 텍스트가 입력되는 문자 임베딩(character embedding) 생성 유닛(613), 문자 임베딩 생성 유닛(613)에 연결되는 3개의 컨볼루션 레이어들(3 conv layers, 615), 및 3개의 컨볼루션 레이어들(615)에 연결되는 양방향 LSTM(bidirectional long short term memory, 617)를 구비할 수 있다. 3개의 컨볼루션 레이어들은 3개의 컨볼루션 신경망(convolutional neural network) 레이어들에 해당하고, '인코더 컨볼루션 레이어들'로 지칭될 수 있다.

[0099] 또한, 스펙트럼 예측기(600)는 인코더(610)의 입력단에 입력 텍스트(input text)를 전처리하는 전처리부(611)를 더 구비할 수 있다. 전처리부(611)는 입력 텍스트를 음절 단위로 분리하고, 분리된 음절을 원핫 인코딩(one-hot encoding)을 통해 정수로 표현하도록 구성될 수 있다. 한편, 이러한 전처리부(611)는, 독립적인 전처리 수단이나 이에 상응하는 기능을 수행하는 독립적인 구성부에 의해 입력 텍스트를 별도로 처리하는 경우에, 생략될 수 있다. 다만, 이러한 전처리부(611)는 스펙트럼 예측기(600)에는 포함되지 않으나 TTS 모델에는 포함되도록 구성될 수 있다.

[0100] 인코더(610)에서, 문자 임베딩 생성 유닛(613)은 전처리부(611)로부터 받은 정수 시퀀스를 매트릭스 형태로 변환한다. 인코더 컨볼루션 레이어들(615)은 매트릭스 형태의 정보를 축약한다. 그리고 양방향 LSTM(617)은 축약된 매트릭스 형태의 정보를 인코더 특징 정보로 변환한다. 여기서 인코더 특징 정보는 하나의 고정된 크기로 압축된 컨텍스트 벡터를 포함할 수 있다.

[0101] 어텐션(620)은 멜겐(MelGAN, 700a)을 통해 출력될 음성 발음이 입력 텍스트의 순차적인 순서대로 진행되도록 디코더(630)의 타임-스텝에 따라 어텐션 얼라인 정보(attention alignment information, AAI)를 인코더 특징 정보에 추가한다. 즉, 어텐션(620)은 매 시점 디코더(630)에서 사용할 정보를 인코더(610)에서 추출하여 정렬(alignment) 혹은 매핑(mapping)하는 과정을 의미할 수 있다. 좀더 구체적으로, 어텐션(620)은 소정 시점마다 디코더(630)에서 집중해서 사용할 정보를 인코더(610)에서 추출하여 할당하고, 매 시점에서 사용하는 정보는 이전 시점의 어텐션 얼라인 정보를 사용하도록 구성될 수 있다. 이러한 구성에 의하면, 어텐션(620)의 얼라인 그

래프는 TTS 모델이 학습 중일 때 우상향으로 연속되어 나갈 수 있다. 이러한 어텐션(620)은 로케이션 센서티브 어텐션(location sensitive attention)을 사용하여 구현될 수 있다.

[0102] 일례로, 로케이션 센서티브 어텐션은 어텐션 점수(attention score)의 계산에서, 이전 시점의 어텐션와 인코더 정보의 가중 합에 아크탄젠트(tanh)를 적용한 후 가중치를 곱하여 계산한다. 여기서, 어텐션은 어텐션 점수에 소프트맥스(softmax)를 적용해 0 내지 1의 범위로 정규화한 것으로 인코더 정보를 얼마나 할당할지 결정하는 정보에 해당된다. 어텐션 추출 결과는 컨텍스트(context)로 지칭되고, 해당 시점까지의 어텐션과 인코더의 곱의 총합에 해당한다.

[0103] 디코더(630)는 어텐션(620)을 통해 어텐션 정보인 얼라인 특징(alignment feature)와 이전 타임-스텝에서 생성된 멜 스펙트로그램을 이용하여 다음 타임-스텝의 멜 스펙트로그램을 생성한다. 생성되는 멜 스펙트로그램에는 통합 스타일 임베딩에 의한 음색과 운율 정보가 포함된다. 이전 타임-스템은 이전 시점에, 다음 타임-스텝은 다음 시점 또는 현재 시점에 각각 대응될 수 있다.

[0104] 디코더(630)는 프리넷(pre-Net, 631), 두 개의 디코더 LSTM(632), 제1 완전연결층(fully connected layer, 633), 제2 완결연결층(634) 및 포스트넷(post-Net, 635)를 구비한다. 프리넷(631)은 2개 레이어로 구성된 프리넷(2 layer pre-Net)일 수 있고, 각 디코더 LSTM(632)은 단방향 LSTM(long short term memory)일 수 있고, 완전연결층은 선형 프로젝션(linear projection)일 수 있으며, 포스트넷(635)은 5개의 컨볼루션 레이어들로 구성된 포스트넷(5 convolution layer post-Net)일 수 있다.

[0105] 디코더(630)에서, 프리넷(631)은 이전 시점의 멜 벡터에서 정보를 추출한다. 디코더 LSTM(632)은 어텐션(620)과 프리넷(631)의 작동 결과를 이용하여 현재 시점의 정보를 추출한다. 제1 완전연결층(633)은 현재 시점의 정보를 이용하여 멜 스펙트로그램(Mel spectrogram)을 생성한다. 제2 완전연결층(634)는 시그모이드(sigmoid)와 결합되어 디코더 LSTM(632)으로부터 나오는 정보를 토대로 현재 시점의 종료 확률을 계산하여 종료 토큰(stop token)을 출력한다. 종료 확률은 0 내지 1의 범위에서 산출될 수 있다. 그리고, 포스트넷(635)은 5개의 1차원(1D) 컨볼루션 레이어와 배치 정규화(batch normalization)를 통해 멜 스펙트로그램의 품질을 향상시킨다. 포스트넷(635)은 디코더(630)을 통해 멜 스펙트로그램이 모두 생성된 후 적용될 수 있고, 생성된 멜 스펙트로그램의 품질을 잔여 커넥션을 이용하여 스무딩(smoothing)하게 보정할 수 있다.

[0106] 전술한 스펙트럼 예측기(600)에서, 손실 함수의 전체 손실은 MSE(mean squared error)와 BCE(binary cross entropy)의 합으로 계산될 수 있다. MSE는 원본 멜 스펙트로그램과 추정 멜 스펙트로그램 간의 차이를 나타내고, BCE는 실제 종료 확률과 추정 종료 확률 간의 차이를 나타낸다. 원본 멜 스펙트로그램은 타겟 멜 스펙트로그램에 대응될 수 있다.

[0107] 멜겐(MelGAN, 700a)는 멜 스펙트로그램을 입력받고 파형(waveform) 신호 또는 음성(audio) 신호를 생성한다. 멜겐(700a)은 생성자(generator)와 판별자(discriminator)로 구성된 GAN 기반 보코더로서 1차원 컨볼루션(conv1d) 여러 층으로 이루어진 모델이다.

[0108] 멜겐(700a)의 생성자는 여러 층의 conv1d와 잔여 스택(residual stack)에 의한 구조로 이루어진다. 잔여 스택 내부에는 또 다른 conv1d가 있고, 잔여 커넥션을 3번 거치도록 구성된다. 생성자는 입력으로 배치 단위의 멜 스펙트로그램을 받는다. 따라서 입력 텐서의 크기는 [배치 사이즈, 80, 프레임 길이]가 될 수 있다. 여기서 80은 멜 스펙트로그램의 차수가 되고, 프레임 길이는 어느 구간만큼 입력할지 사용자가 정하는 변수이다. 출력은 오디오 신호과 되고 그 크기는 [배치 사이즈, 1, 프레임 길이×홉 사이즈)가 된다.

[0109] 멜겐(700a)의 판별자는 3개의 다중 스케일(multi-scale) 구조로 이루어진다. 하나의 다중 스케일에는 6개의 특징맵(feature map)과 1개의 출력으로 총 7개의 출력을 가진다. 판별자 블록은 총 7개의 conv1d로 구성될 수 있다.

[0110] 멜겐(700a)의 마지막 출력은 기본적으로 판별자 출력이고, 나머지 출력은 손실(loss)로 사용될 수 있다.

[0112] 도 5는 본 발명의 다른 실시예에 따른 음성합성 시스템의 전체 구성도이다. 도 6은 도 5의 음성합성 시스템에 채용할 수 있는 감정 모듈을 설명하기 위한 블록도이다. 그리고 도 7은 도 6의 감정 모듈에 채용할 수 있는 레퍼런스 인코더에 대한 상세 구성도이다

[0113] 도 5를 참조하면, 음성합성 시스템은 운율 모듈(100), 음색 모듈(200), 감정 모듈(300), 스펙트럼 예측기(spectrum predictor, 600) 및 멜겐(MelGAN, 700a)를 포함하고, 운율 모듈(100)의 제1 스타일 임베딩(style embedding)과 음색 모듈(200)의 제2 스타일 임베딩과 감정 모듈(300)의 제3 스타일 임베딩을 합한 통합 스타일

임베딩을 적용하여 입력 텍스트(input text)에 특정 운율, 음색 및 감정이 부가된 파형 샘플(waveform samples)을 출력할 수 있다. 스펙트럼 예측기(600) 및 멜겐(MelGAN, 700a)은 음성합성(text to speech, TTS) 모델(500)을 구성한다.

[0114] 운율 모듈(100), 음색 모듈(200), 스펙트럼 예측기(600) 및 멜겐(MelGAN, 700a)은 도 4를 참조하여 앞서 설명한 음성합성 시스템의 대응 구성요소와 실질적으로 동일하므로 이들에 대한 상세 설명은 생략하기로 한다.

[0115] 감정 모듈(300)은 STFT 유닛(310), 제3 레퍼런스 인코더(320) 및 제5 스타일 토큰 레이어(style token 5, 350)을 구비한다. STFT 유닛(310)은, 참조 오디오를 별도로 가공한 스펙트로그램이나 멜 스펙트로그램을 제3 레퍼런스 인코더(320)의 입력 데이터로 준비하는 경우에, 생략될 수 있다. 제5 스타일 토큰 레이어(350)는 제5 GST 레이어에 대응된다.

[0116] 감정 모듈(300)에서, STFT 유닛(310)은 참조 오디오를 푸리에 변환에 의해 변환한 스펙트로그램이나 멜 스펙트로그램을 제3 레퍼런스 인코더(320)의 입력(input)으로 제공할 수 있다. 제3 레퍼런스 인코더(320)는 기본주파수가 아닌 스펙트로그램이나 멜 스펙트로그램을 입력받아 감정 임베딩(emotional embedding)을 생성한다.

[0117] 제5 GST 레이어(350)를 통과한 감정 임베딩은 제3 스타일 임베딩으로 출력된다. 즉, 제5 스타일 토큰 레이어(350)는 감정 임베딩을 학습하여 제3 스타일 임베딩을 생성할 수 있다. 이러한 제5 스타일 토큰 레이어(350)는 다중헤드 어텐션(multi-head attention, 352) 및 복수의 토큰들(354)을 구비한다. 복수의 토큰들(354)은 제1 토큰(token 1) 내지 제k 토큰(token k)을 포함할 수 있고, 각 토큰은 스타일 토큰으로 지칭될 수 있다. 감정 임베딩의 효과적을 학습을 위해 상기의 k는 7인 것이 바람직하다. 즉, 본 실시예의 제5 GST 레이어(350)는 7개로 분류된 감정 스타일에 대응하도록 7개의 스타일 토큰들을 구비할 수 있다.

[0118] 감정 스타일에 포함되는 7개의 감정은 분노(Anger), 실망(Disappointment), 두려움(Fear), 놀람(Surprise), 슬픔(Sad), 평온함(Neutral) 및 행복함(Happy)이고, 이를 MOS(mean opinion score) 점수와 함께 정리하면 다음의 표 1과 같다.

표 1

[0119]

	MOS Scores
Anger	3.87
Disappointment	4.14
Fear	4.85
Surprise	3.71
Sad	4.71
Neutral	4.71
Happy	3.57

[0120] 전술한 제5 스타일 토큰 레이어(350)는, 도 6에 도시한 바와 같이, 다중헤드 어텐션(352)에 의해 제3 레퍼런스 인코더(320)에서 출력된 감정 임베딩과 7개의 스타일 토큰들 각각과의 유사도가 측정될 때(360 참조), 이 스타일 토큰들(244)의 가중 합으로 스타일 임베딩(제3 스타일 임베딩)을 생성할 수 있다.

[0121] 감정 모듈(300)에서 생성된 제3 스타일 임베딩은 음율 모듈(100)에서 총 3개의 스타일 토큰 레이어들을 통과한 제1 스타일 임베딩과 음색 모듈(200)의 제2 스타일 임베딩과 합쳐지고(concatenate) 음성합성 시스템의 스펙트럼 예측기(600)의 인코더(encoder, 610))의 출력과 결합되어 스펙트럼 예측기(600)의 어텐션(620)에 입력될 수 있다.

[0122] 한편, 도 7을 참조하여 전술한 제3 레퍼런스 인코더(320)를 좀더 구체적으로 설명하면, 입력되는 참조 스펙트로그램(reference spectrogram)을 변환하여 감정 임베딩(emotional embedding)을 출력하기 위해, 제3 레퍼런스 인코더(320)는 3개의 합성곱 신경망(three convolutional neural networks, 3 CNN)(322), 3개의 합성곱 신경망(322)에 연결되는 2개의 잔여 블록들(2 residual blocks, 324), 및 2개의 잔여 블록들(324)에 연결되는 수정 알렉스넷(modified AlexNet, 326)으로 구성될 수 있다. 참조 스펙트로그램은 참조 오디오의 입력을 푸리에 변환을 통해 변환한 스펙트로그램을 지칭한다. 참조 스펙트로그램은 참조 멜 스펙트로그램으로 대체될 수 있다.

[0123] 이러한 제3 레퍼런스 인코더(320)를 포함한 감정 모듈에 결합하는 TTS 모델은, 참조 오디오의 감정 표현을 음성 합성에 적용하기 위해, 스타일 트랜스퍼 및 스타일 트랜스퍼의 강도를 조절하도록 구성되고, 리컨스트럭션 로스(reconstruction loss)만으로 학습을 진행할 수 있다.

[0125] 도 8은 본 발명의 또 다른 실시예에 따른 음성합성 시스템에 대한 개략적인 구성도이고, 도 9는 도 8의 음성합성 시스템에 적용될 수 있는 주요 동작을 설명하기 위한 블록도이다.

[0126] 도 8을 참조하면, 음성합성 시스템(1000)은, 프로세서(processor, 1100) 및 메모리(momory, 1200)를 포함하여 구성될 수 있다. 또한, 음성합성 시스템(100)은 송수신 장치(transceiver, 1300)를 더 포함하거나, 저장 장치(1400)를 더 포함하거나, 입력 인터페이스 장치(1500) 및 출력 인터페이스 장치(1500)를 더 포함하도록 구성될 수 있다.

[0127] 음성합성 시스템(1000)에서, 프로세서(1100)는 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 발명의 실시예들에 따른 방법들이 수행되는 전용의 프로세서를 의미할 수 있다.

[0128] 메모리(1200) 또는 저장 장치(1400)는 프로세서(1100)에 의해 실행되는 적어도 하나의 명령을 저장할 수 있다. 적어도 하나의 명령은, 음율 모듈을 구현하거나 동작시키기 위한 제1 명령, 음색 모듈을 구현하거나 동작시키기 위한 제2 명령, 감정 모듈을 구현하거나 동작시키기 위한 제3 명령, TTS 모델을 구현하거나 동작시키기 위한 제5 명령을 포함할 수 있다.

[0129] 또한, 적어도 하나의 명령은, 입력되는 참조 오디오에서 기본주파수를 추출하는 명령, 기본주파수를 레퍼런스 인코더의 입력으로 전달하는 명령, 기본주파수를 인코딩하여 운율 임베딩을 생성하는 명령, 운율 임베딩으로부터 제1 스타일 임베딩을 생성하는 명령, 참조 오디오를 푸리에 변환에 의해 참조 멜 스펙트로그램으로 변환하는 명령, 참조 멜 스펙트로그램을 인코딩하여 스피커 임베딩을 생성하는 명령, 스피커 임베딩으로부터 제2 스타일 임베딩을 생성하는 명령, 제1 스타일 임베딩과 제2 스타일 임베딩을 합한 통합 스타일 임베딩을 음성합성(text to speech, TTS) 모델의 인코더의 출력과 합쳐 어텐션에 입력하는 명령, TTS 모델의 어텐션과 디코더를 통해 멜 스펙트로그램을 만드는 명령, TTS 모델의 보코더를 통해 입력 텍스트로부터 통합 스타일 임베딩에 의한 음색(tones)과 운율(prosody)이 결합된 음성합성 오디오를 생성하는 명령을 포함할 수 있다.

[0130] 전술한 메모리(1200) 및 저장 장치(1400) 각각은 휘발성 저장 매체 및 비휘발성 저장 매체 중에서 적어도 하나로 구성될 수 있다. 예를 들어, 메모리(1200) 또는 저장 장치(1400)는 읽기 전용 메모리(read only memory, ROM) 및 랜덤 액세스 메모리(random access memory, RAM) 중에서 적어도 하나로 구성될 수 있다.

[0131] 송수신 장치(1300)는 무선 네트워크, 유선 네트워크, 위성 네트워크 또는 이들의 조합을 통해 외부 장치와의 통신을 지원하는 적어도 하나의 서브통신시스템을 포함할 수 있다.

[0132] 입력 인터페이스 장치(1500)는 키보드, 마이크, 터치패드, 터치스크린 등의 입력 수단들에서 선택되는 적어도 하나와 적어도 하나의 입력 수단을 통해 입력되는 신호를 기저장된 명령과 매핑하거나 처리하는 입력 신호 처리부를 포함할 수 있다.

[0133] 출력 인터페이스 장치(1600)는 프로세서(1100)의 제어에 따라 출력되는 신호를 기저장된 신호 형태나 레벨로 매핑하거나 처리하는 출력 신호 처리부와, 출력 신호 처리부의 신호에 따라 진동, 빛 등의 형태로 신호나 정보를 출력하는 적어도 하나의 출력 수단을 포함할 수 있다. 출력 신호 처리부는 배선계통의 네트워크 토폴로지나 전압 상태추정 결과를 이미지, 음성 또는 이들의 조합 형태로 생성할 수 있다. 그리고 적어도 하나의 출력 수단은 스피커, 디스플레이 장치, 프린터, 광 출력 장치, 진동 출력 장치 등을 포함할 수 있다.

[0134] 전술한 음성합성 시스템(1000)은, 예를 들어, 데스크탑 컴퓨터(desktop computer), 랩탑 컴퓨터(laptop computer), 노트북(notebook), 스마트폰(smart phone), 태블릿 PC(tablet PC), 모바일폰(mobile phone), 스마트 워치(smart watch), 스마트 글래스(smart glass), e-book 리더기, PMP(portable multimedia player), 휴대용 게임기, 네비게이션(navigation) 장치, 디지털 카메라(digital camera), DMB(digital multimedia broadcasting) 재생기, 디지털 음성 녹음기(digital audio recorder), 디지털 음성 재생기(digital audio player), 디지털 동영상 녹화기(digital video recorder), 디지털 동영상 재생기(digital video player), PDA(Personal Digital Assistant), 네트워크 서버, 웹서버, 메일서버, 특정 서비스를 위한 서비스 서버 등에 일체로 결합되거나 탑재될 수 있다.

[0135] 한편, 음성합성 시스템(1000)은, 도 9에 도시한 바와 같이, 학습 관리를 위한 제1 모듈(1700)과 스타일 복제

TTS 서비스를 위한 제2 모듈(180)을 더 구비할 수 있다.

[0136] 제1 모듈(1700)은 원하는 음색과 운율을 가진 제1 오디오 파형이나 원하는 음색과 운율과 감정을 가진 제2 오디오 파형을 그라운드 트루(ground truth, GT)의 참조 오디오로 사용하여, 운율 모듈의 계층형 GST와 음색 모듈의 제4 GST 레이어를 훈련시키거나, 감정 모듈의 제5 GST 레이어를 더 훈련시키도록 구성될 수 있다.

[0137] 또한, 음성합성 시스템(1000)은, 제2 모듈(1800)에 의해 참조 오디오의 음색과 운율 스타일을 복제한 음성합성(TTS) 서비스를 제공하거나, 참조 오디오의 음색, 운율 및 감정 스타일을 복제한 음성합성 서비스를 제공하도록 구성될 수 있다.

[0138] 이러한 제1 모듈(1700) 및/또는 제2 모듈(1800)은 프로그램 명령이나 소프트웨어 모듈로 구현되어 메모리(1200)이나 저장 장치(1400)에 저장되고, 프로세서(1100)에 의해 실행되도록 구성될 수 있다. 물론, 구현에 따라서, 제1 모듈(1700) 및/또는 제2 모듈(1800)의 적어도 일부는 해당 기능의 적어도 일부를 위한 하드웨어 구성을 포함하도록 구성될 수 있다.

[0139] 전술한 바와 같이, 음성합성 시스템(100)은 TTS 모델과 함께 운율 모듈 및 음색 모듈을 포함하거나 선택적으로 감정 모듈을 더 포함하도록 구성되고, 이 모듈들 각각은 전역 스타일 토큰을 기반으로 원하는 음색, 운율, 감정 또는 이들 조합의 스타일을 포함하고 있는 참조 오디오만 있다면 참조 오디오의 스타일대로 음성합성을 구현할 수 있는 장점이 있다.

[0140] 본 발명의 실시예에 따른 방법의 동작은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

[0141] 또한, 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

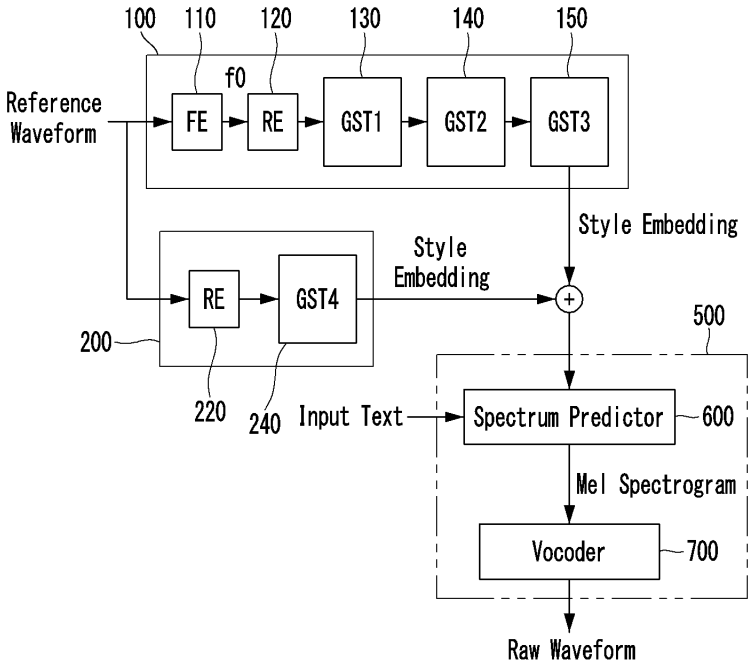
[0142] 본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시예에서, 가장 중요한 방법 단계들의 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0143] 실시예들에서, 프로그램 가능한 로직 장치 예컨대, 필드 프로그래머블 게이트 어레이가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시예들에서, 필드 프로그래머블 게이트 어레이는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

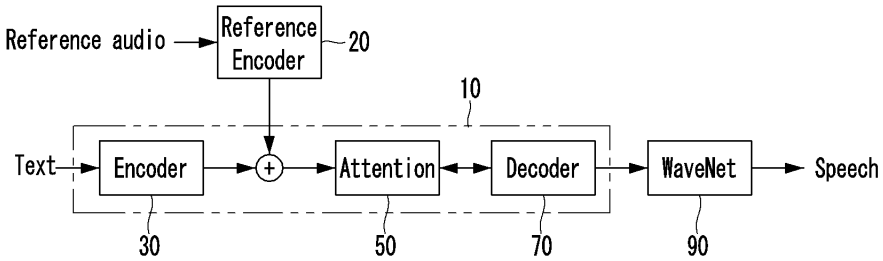
[0144] 이상 본 발명의 바람직한 실시예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

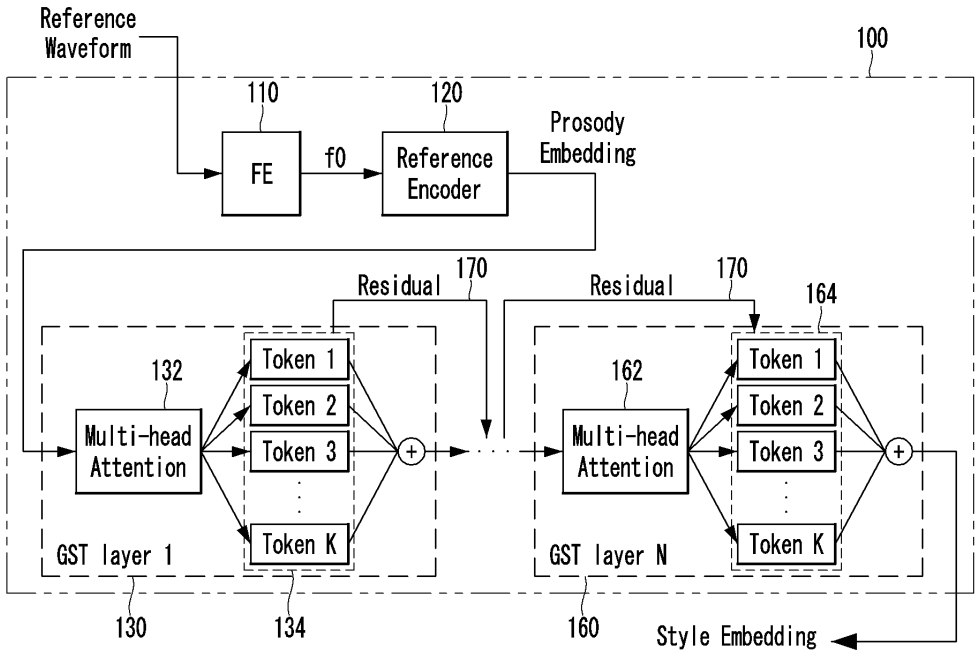
도면1


100
110
120
130
140
150
f0
Reference Waveform
FE
RE
GST1
GST2
GST3
Style Embedding
Style Embedding
RE
GST4
200
220
240
Input Text
500
Spectrum Predictor
600
Mel Spectrogram
Vocoder
700
Raw Waveform

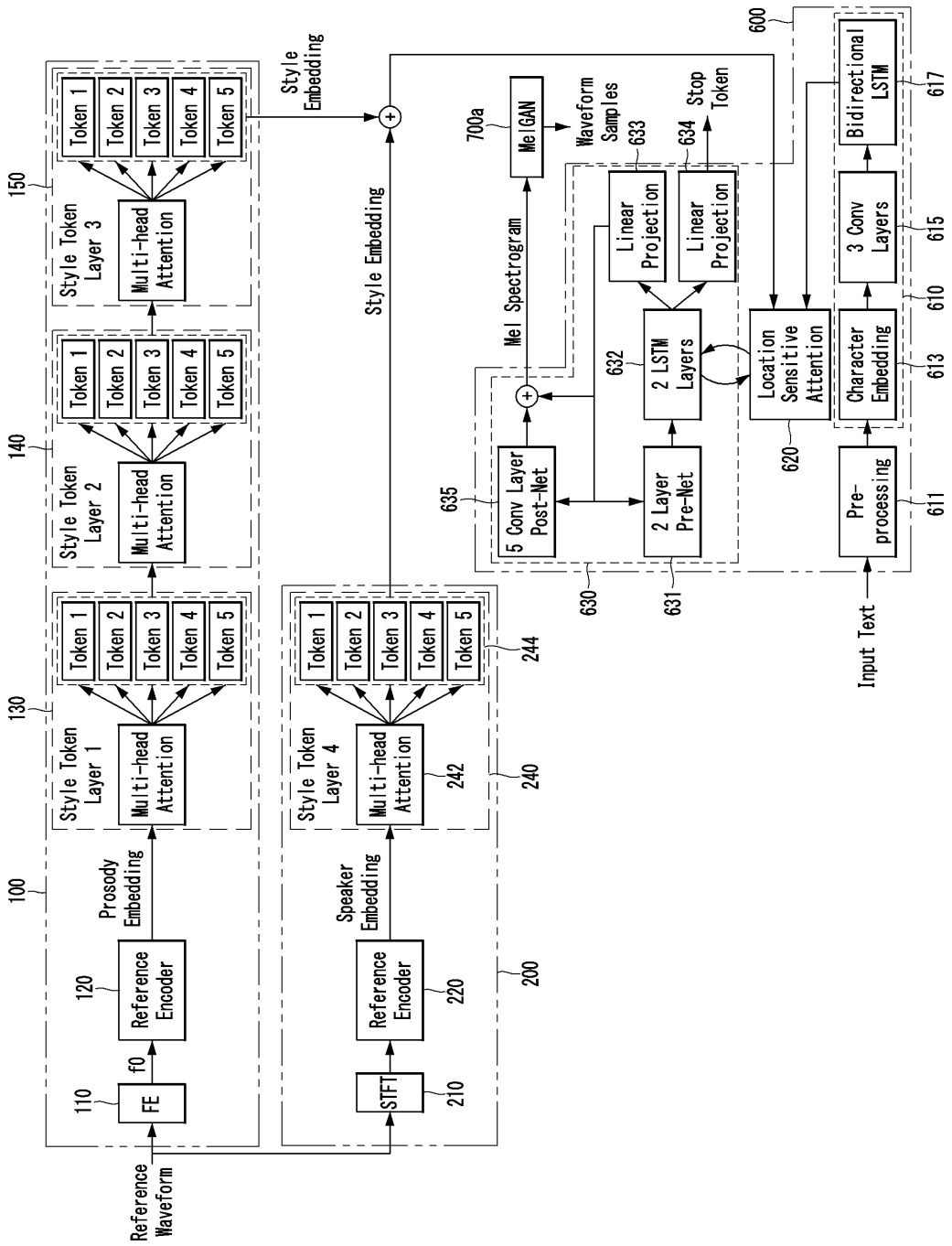
도면2


Reference audio
Reference Encoder
20
10
Text
Encoder
Attention
Decoder
WaveNet
Speech
30
50
70
90

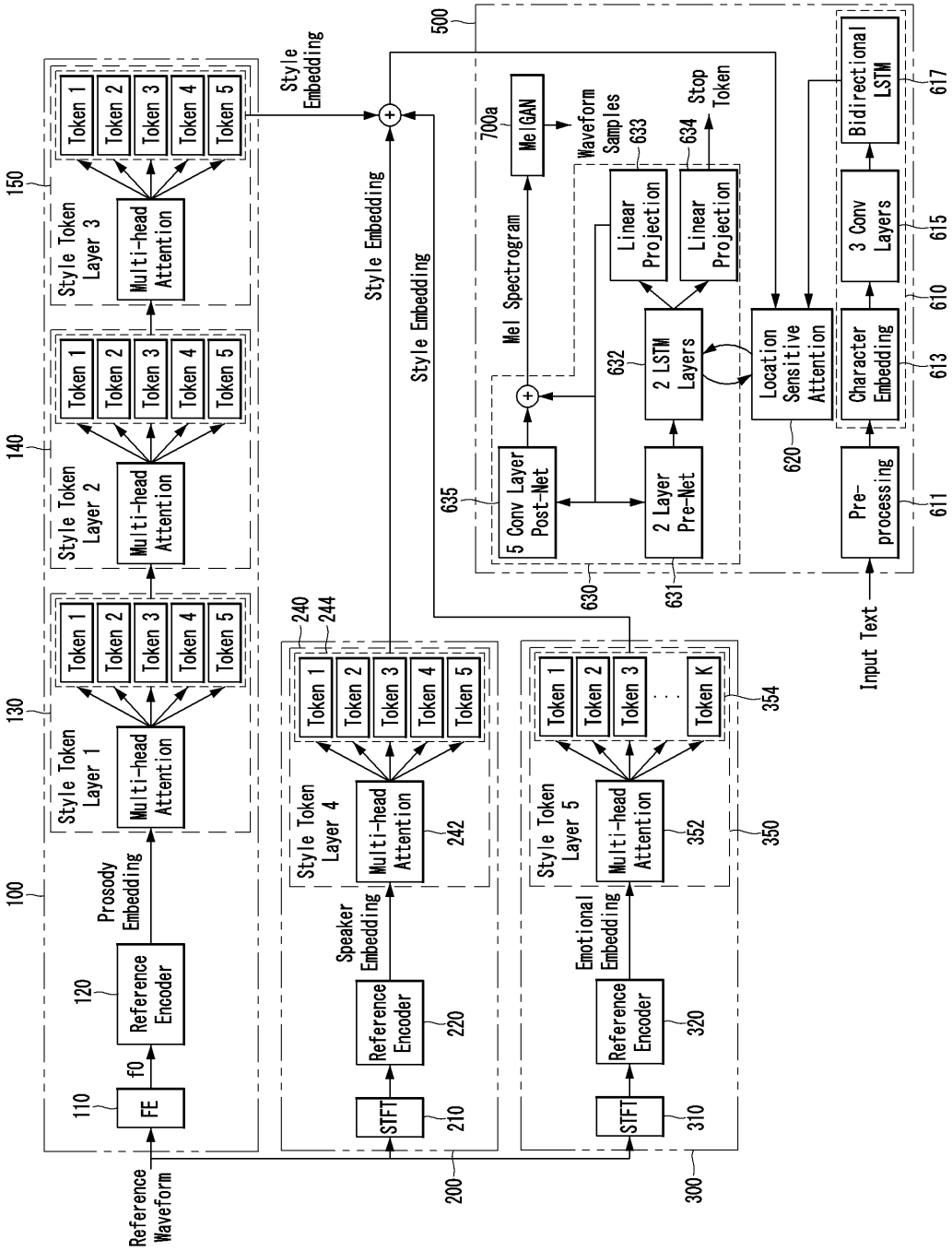
도면3


Reference Waveform
100
110
120
FE
f0
Reference Encoder
Prosody Embedding
Residual
170
Residual
170
164
132
Multi-head Attention
Token 1
Token 2
Token 3
Token K
GST layer 1
130
134
162
Multi-head Attention
Token 1
Token 2
Token 3
Token K
GST layer N
160
Style Embedding

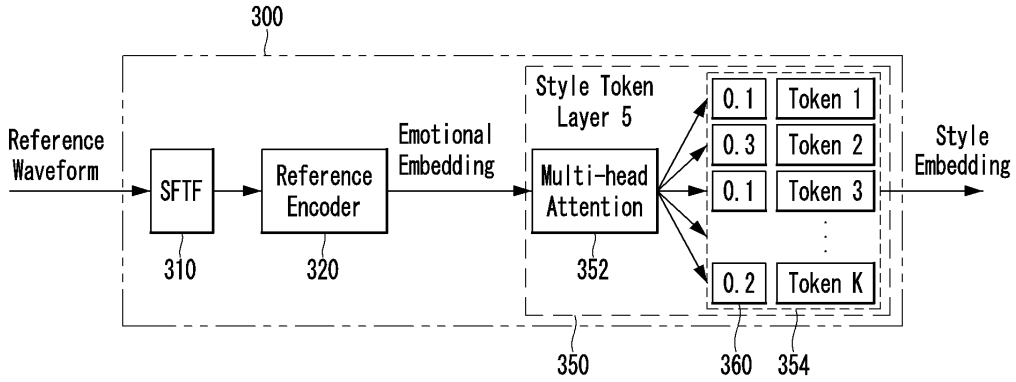
도면4


Reference Waveform
110
FE
f0
120
Reference Encoder
Prosody Embedding
100
130
Style Token Layer 1
Multi-head Attention
140
Style Token Layer 2
Multi-head Attention
150
Style Token Layer 3
Multi-head Attention
Token 1
Token 2
Token 3
Token 4
Token 5
Style Embedding
STFT
210
Reference Encoder
220
Speaker Embedding
200
240
Style Token Layer 4
Multi-head Attention
242
244
Style Embedding
Input Text
600
Pre-processing
611
Character Embedding
613
3 Conv Layers
615
Bidirectional LSTM
617
610
Location Sensitive Attention
620
2 LSTM Layers
632
2 Layer Pre-Net
631
5 Conv Layer Post-Net
635
630
Mel Spectrogram
Linear Projection
633
Linear Projection
634
Stop Token
700a
MelGAN
Waveform Samples

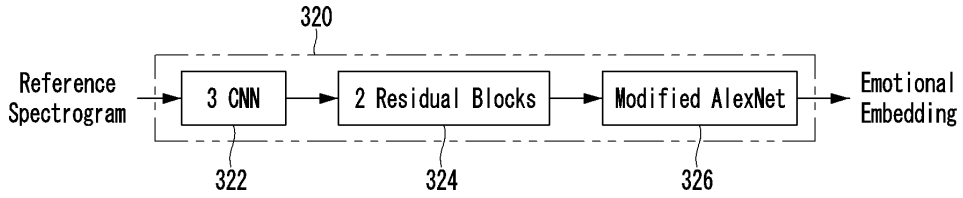
도면5


Reference Waveform
200
300
100
110
FE
f0
120
Reference Encoder
Prosody Embedding
130
Style Token Layer 1
Multi-head Attention
Token 1
Token 2
Token 3
Token 4
Token 5
140
Style Token Layer 2
150
Style Token Layer 3
Style Embedding
240
244
Style Token Layer 4
242
Speaker Embedding
STFT
210
220
Reference Encoder
350
Style Token Layer 5
352
Emotional Embedding
310
320
354
Token K
500
700a
MelGAN
Waveform Samples
633
Mel Spectrogram
630
631
635
5 Conv Layer Post-Net
2 Layer Pre-Net
632
2 LSTM Layers
Linear Projection
634
Stop Token
620
Location Sensitive Attention
Input Text
611
Pre-processing
613
Character Embedding
610
3 Conv Layers
615
Bidirectional LSTM
617

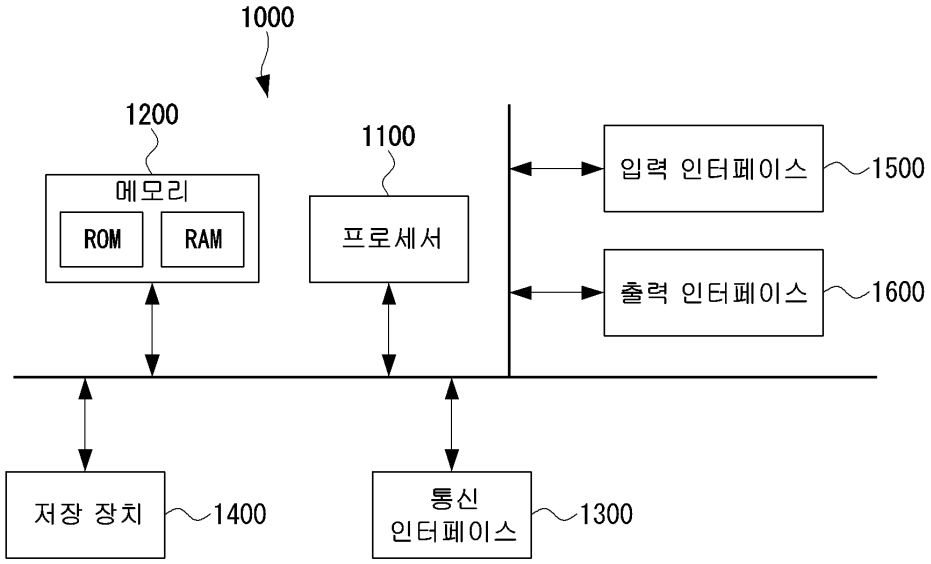
도면6


300
Reference Waveform
SFTF
310
Reference Encoder
320
Emotional Embedding
Style Token Layer 5
Multi-head Attention
352
0.1
0.3
0.1
0.2
Token 1
Token 2
Token 3
Token K
Style Embedding
350
360
354

도면7


320
Reference Spectrogram
3 CNN
2 Residual Blocks
Modified AlexNet
Emotional Embedding
322
324
326

도면8


1000
1200
메모리
ROM
RAM
1100
프로세서
입력 인터페이스
1500
출력 인터페이스
1600
저장 장치
1400
통신 인터페이스
1300

도면9