(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0029686
(43) 공개일자 2025년03월05일

(51) 국제특허분류(Int. Cl.)
G10L 13/10 (2013.01) *G06F 40/279* (2020.01)
G06N 3/044 (2023.01) *G06N 3/0464* (2023.01)
G06N 3/0895 (2023.01) *G10L 13/02* (2006.01)
G10L 17/02 (2013.01) *G10L 21/0208* (2013.01)
(52) CPC특허분류
G10L 13/10 (2013.01)
G06F 40/279 (2020.01)
(21) 출원번호 10-2023-0132536
(22) 출원일자 2023년10월05일
심사청구일자 2023년10월05일
(30) 우선권주장
1020230110815 2023년08월23일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이성환
서울특별시 강남구 언주로30길 13, B-2506(도곡동, 대림아크로빌, 대림아크로텔)
오형석
서울특별시 성북구 고려대로27길 47 (안암동5가)
이상훈
서울특별시 성북구 인촌로27길 34-8 (안암동5가)
(74) 대리인
윤귀상

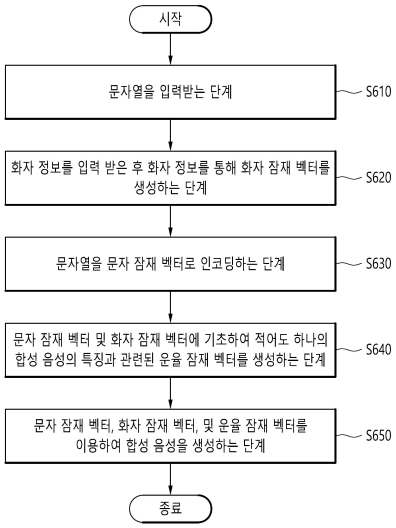
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **확산 기반 잠재 운율 생성 및 운율 조건부 적대적 훈련을 통한 표현적 음성 합성을 수행하는 방법 및 장치**

(57) 요 약

음성 합성 방법 및 음성 합성 장치가 개시된다. 본 개시의 일 실시예에 따른, 음성 합성 장치는 문자열을 입력받는 문자 입력부; 화자 정보를 입력 받은 후 상기 화자 정보를 통해 화자 잠재 벡터를 생성하는 화자 입력부; 상기 문자열을 문자 잠재 벡터로 인코딩하는 문자 인코더부; 상기 문자 잠재 벡터 및 상기 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성하도록 학습된 운율 잠재 벡터 예측부; 및 상기 문자 잠재 벡터, 상기 화자 잠재 벡터, 및 상기 운율 잠재 벡터를 이용하여 합성 음성을 생성하는 음성 합성부를 포함할 수 있다.

대 표 도 - 도6



(52) CPC특허분류
G06N 3/044 (2023.01)
G06N 3/0464 (2023.01)
G06N 3/0895 (2023.01)
G10L 13/02 (2013.01)
G10L 17/02 (2013.01)
G10L 21/0208 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193973
과제번호 2019-0-00079-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 인공지능대학원지원(고려대학교)
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711193622
과제번호 2021-0-02068-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 AI혁신허브 사업
연구과제명 인공지능 혁신 허브 연구 개발
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

문자열을 입력받는 문자 입력부;

화자 정보를 입력 받은 후 상기 화자 정보를 통해 화자 잠재 벡터를 생성하는 화자 입력부;

상기 문자열을 문자 잠재 벡터로 인코딩하는 문자 인코더부;

상기 문자 잠재 벡터 및 상기 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성하도록 학습된 운율 잠재 벡터 예측부; 및

상기 문자 잠재 벡터, 상기 화자 잠재 벡터, 및 상기 운율 잠재 벡터를 이용하여 합성 음성을 생성하는 음성 합성부를 포함하는, 음성 합성 장치.

청구항 2

제1항에 있어서,

참조 음성을 입력받는 음성 입력부;

상기 참조 음성을 참조 음성 운율 잠재 벡터로 인코딩하는 참조 음성-운율 벡터 인코더부; 및

상기 합성 음성 및 상기 참조 음성 운율 잠재 벡터에 기초하여 상기 합성 음성이 비-합성 음성인지 여부를 판별하는 운율 조건부 판별부를 더 포함하는, 음성 합성 장치.

청구항 3

제2항에 있어서,

상기 문자 인코더부 및 상기 합성 음성부는, 상기 운율 조건 판별부에 의해 적대적 조건 훈련 방식을 통해 상기 합성 음성이 비-합성 음성으로 판별되도록 학습되는, 음성 합성 장치.

청구항 4

제1항에 있어서,

상기 운율 잠재 벡터 예측부는, 노이즈 제거 확산 확률 모델을 포함하고,

상기 노이즈 제거 확산 확률 모델은:

전진 과정(forward process)를 통해 학습 운율 잠재 벡터에 대해 노이즈를 부가하고,

후진 과정(reverse process)를 통해 상기 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하도록 학습되는, 음성 합성 장치.

청구항 5

제1항에 있어서,

상기 학습 운율 잠재 벡터에 대해 노이즈를 부가하는 노이즈 입력부; 및

상기 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하는 노이즈 제거부를 포함하고,

상기 노이즈 입력부 또는 상기 노이즈 제거부 중의 적어도 하나는, 합성곱 신경망(convolutional neural network, CNN), 순환 신경망(recurrent neural network, RNN), 장-단기 기억(long short term memory, LSTM), GRU(gated-recurrent unit) 또는 트랜스포머(transformer) 중의 적어도 하나를 포함하는, 음성 합성 장치.

청구항 6

문자열을 입력받는 단계;

화자 정보를 입력 받은 후 상기 화자 정보를 통해 화자 잠재 벡터를 생성하는 단계;

상기 문자열을 문자 잠재 벡터로 인코딩하는 단계;

상기 문자 잠재 벡터 및 상기 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성하는 단계; 및

상기 문자 잠재 벡터, 상기 화자 잠재 벡터, 및 상기 운율 잠재 벡터를 이용하여 합성 음성을 생성하는 단계를 포함하는, 음성 합성 방법.

청구항 7

제6항에 있어서,

참조 음성을 입력받는 단계;

상기 참조 음성을 참조 음성 운율 잠재 벡터로 인코딩하는 단계; 및

상기 합성 음성 및 상기 참조 음성 운율 잠재 벡터에 기초하여 상기 합성 음성이 비-합성 음성인지 여부를 판별하는 단계를 더 포함하는, 음성 합성 방법.

청구항 8

제7항에 있어서,

상기 문자열을 상기 문자 잠재 벡터로 인코딩하는 문자 인코더부 및 상기 합성 음성을 생성하는 합성 음성부는, 적대적 조건 훈련 방식을 통해 상기 합성 음성이 비-합성 음성으로 판별되도록 학습되는, 음성 합성 방법.

청구항 9

제6항에 있어서,

상기 합성 음성이 비-합성 음성인지 여부를 판별하는 단계는 노이즈 제거 확산 확률 모델을 통해 수행되고,

상기 노이즈 제거 확산 확률 모델은:

전진 과정(forward process)를 통해 학습 운율 잠재 벡터에 대해 노이즈를 부가하고,

후진 과정(reverse process)를 통해 상기 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하도록 학습되는, 음성 합성 방법.

청구항 10

제6항에 있어서,

상기 학습 운율 잠재 벡터에 대해 노이즈를 부가하는 노이즈 입력부 및 상기 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하는 노이즈 제거부 중의 적어도 하나는, 합성곱 신경망(convolutional neural network, CNN), 순환 신경망(recurrent neural network, RNN), 장-단기 기억(long short term memory, LSTM), GRU(gated-recurrent unit) 또는 트랜스포머(transformer) 중의 적어도 하나를 포함하는, 음성 합성 방법.

발명의 설명

기 술 분 야

[0001] 본 개시는 잠재 운율 생성을 통한 표현적 음성 합성 방법에 관한 것이며, 보다 상세하게는 확률적 확산 모델로 표현적이고 정확한 잠재 운율을 생성하고, 운율 조건부 적대적 훈련으로 표현적인 음성을 생성하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 음성 합성은 기술적으로 텍스트를 음성으로 변환하는 과정을 의미한다. 음성 합성의 배경 기술은 음성 생성을 위한 다양한 기술과 알고리즘의 집합으로 구성되어 있다. 특히, 신경망을 이용한 음성 합성은 최근 음성 합성 기술 발전의 중심으로 부상하고 있다.

[0003] 신경망 모델은 딥러닝 알고리즘을 활용하여 음성 합성에 사용되는 모델을 구축할 수 있다. 이러한 모델은 대규모 데이터 셋에서 훈련되며, 텍스트 입력 및 음성 출력 간의 관계를 학습하여 자연스러운 음성을 생성할 수 있다.

[0004] 표현적인 음성 합성(expressive speech synthesis)은 음성 합성 기술의 한 분야로, 음성에 표현력과 감정을 부여하여 자연스러운 대화를 구현하는 것을 목표로 하며, 운율 모델링(prosody modeling)을 통해 실현될 수 있다.

[0005] 운율 모델링은 참조 기반 방식(reference-based) 및 비-참조 방식(reference-free) 방식으로 구분될 수 있다. 참조 기반 방식은 참조 인코더에 참고하고자 하는 음성을 입력하여 스타일 벡터를 획득하고, 이를 조건으로 음성을 합성하는 방식을 의미한다. 비-참조 방식은 운율의 각 요소(예로, 피치, 억양, 리듬 등)을 직접 모델링하고 이를 조건으로 음성을 합성하는 방식을 의미한다.

[0006] 뿐만 아니라, 운율 요소를 각각 모델링하는 것이 아닌 하나의 잠재 운율 벡터를 모델링하고, 입력 문자 및 화자 정보로부터 잠재 운율 벡터를 자기회기적으로 예측하는 방법 역시 고안되었다.

[0007] 참조 기반 방식은 참조 음성의 운율을 분리 문제를 풀어야 한다는 한계가 존재한다. 비-참조 방식은 운율의 각 요소를 직접 제어할 수 있으나 운율 요소 각각이 서로 의존적이므로, 따로 모델링함에 따라 부자연스러운 음성을 초래할 수 있다는 한계가 존재한다. 운율 잠재 벡터를 자기회기적으로 예측하는 모델은 반복 및 생략과 같은 문제가 발생할 수 있다.

[0008] 노이즈 제거 확산 모듈(denoising diffusion probabilistic model)은 생성 모델의 한 종류를 의미한다. 노이즈 제거 확산 모듈 기반의 동작은 데이터에 노이즈를 더해 붕괴하는 전진 과정(forward process)와 노이즈를 제거해 복원하는 후진 과정(reverse process)으로 이루어진다.

[0009] 노이즈는 데이터 공간에서 확률적으로 확산되어 전파되는 것으로 모델링된다. 이 모델은 여러 단계로 구성되며, 각 단계에서 노이즈가 점진적으로 확산되어 원본 데이터에 가까워질 수 있다. 이러한 확산 과정은 확률론적인 방법을 사용하여 모델을 학습하고, 잡음 추정 및 복원에 활용될 수 있다.

[0010] 적대적 훈련은 생성 모델과 판별 모델 간의 상호 작용을 통해 모델을 향상시키는 기법이다. 이 방법은 생성자와 판별자라는 두 개의 모델을 함께 훈련시키는데, 생성자는 실제 데이터와 유사한 가짜 데이터를 생성하고, 판별자는 생성된 가짜 데이터와 실제 데이터를 구별하도록 설계될 수 있다. 훈련 과정에서 생성자는 판별자를 속이기 위해 더욱 실제와 유사한 데이터를 만들기 위해 학습될 수 있다. 한편, 판별자는 생성된 가짜 데이터와 실제 데이터를 구분하여 정확한 판단을 내리도록 학습할 수 있다.

[0011] 이렇게 생성된 생성자와 판별자 간의 경쟁은 점진적인 개선을 이끌어내며, 생성자는 점차적으로 더 현실적인 결과물을 생성할 수 있으며, 판별자는 더 정확하게 가짜 데이터와 실제 데이터를 구별할 수 있게 된다.

[0012] 적대적 훈련은 생성 모델의 품질 향상을 위한 기법으로 여겨지며, 이미지 생성, 자연어 처리, 음성 합성 등 다양한 분야에서 널리 사용되고 있다. 즉, 적대적 훈련은 생성 모델이 현실적인 결과물을 출력하는데 도움이 되며 데이터 분포를 모델링하는 능력을 향상시킬 수 있다.

발명의 내용

해결하려는 과제

[0013] 본 개시의 목적은 확산 기반 잠재 운율 생성 및 운율 조건부 적대적 훈련을 통한 표현적 음성 합성을 수행하는 방법 및 장치를 제공함에 있다.

[0014] 본 개시의 목적은 운율을 단어 단위 벡터로 모델링하고, 이를 노이즈 제거 확산 확률 모델을 통해 생성하는 방법 및 표현력 있는 음성 합성을 위한 운율 조건부 적대적 훈련 방법 및 장치를 제공함에 있다.

[0015] 본 개시의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 개시의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 개시의 목적 및 장점들은 청구범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

[0016] 본 개시의 일 실시예에 따른 음성 합성 장치는, 문자열을 입력받는 문자 입력부; 화자 정보를 입력 받은 후 상기 화자 정보를 통해 화자 잠재 벡터를 생성하는 화자 입력부; 상기 문자열을 문자 잠재 벡터로 인코딩하는 문자 인코더부; 상기 문자 잠재 벡터 및 상기 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성하도록 학습된 운율 잠재 벡터 예측부; 및 상기 문자 잠재 벡터, 상기 화자 잠재 벡터, 및 상기 운율 잠재 벡터를 이용하여 합성 음성을 생성하는 음성 합성부를 포함할 수 있다.

[0017] 그리고, 음성 합성 장치는, 참조 음성을 입력받는 음성 입력부; 상기 참조 음성을 참조 음성 운율 잠재 벡터로 인코딩하는 참조 음성-운율 벡터 인코더부; 및 상기 합성 음성 및 상기 참조 음성 운율 잠재 벡터에 기초하여 상기 합성 음성이 비-합성 음성인지 여부를 판별하는 운율 조건부 판별부를 더 포함할 수 있다.

[0018] 그리고, 상기 문자 인코더부 및 상기 합성 음성부는, 상기 운율 조건 판별부에 의해 적대적 조건 훈련 방식을 통해 상기 합성 음성이 비-합성 음성으로 판별되도록 학습될 수 있다.

[0019] 그리고, 상기 운율 잠재 벡터 예측부는, 노이즈 제거 확산 확률 모델을 포함하고, 상기 노이즈 제거 확산 확률 모델은: 전진 과정(forward process)를 통해 학습 운율 잠재 벡터에 대해 노이즈를 부가하고, 후진 과정(reverse process)를 통해 상기 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하도록 학습될 수 있다.

[0020] 그리고, 상기 음성 합성 장치는, 상기 학습 운율 잠재 벡터에 대해 노이즈를 부가하는 노이즈 입력부; 및 상기 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하는 노이즈 제거부를 포함하고, 상기 노이즈 입력부 또는 상기 노이즈 제거부 중의 적어도 하나는, 합성곱 신경망(convolutional neural network, CNN), 순환 신경망(recurrent neural network, RNN), 장-단기 기억(long short term memory, LSTM), GRU(gated-recurrent unit) 또는 트랜스포머(transformer) 중의 적어도 하나를 포함할 수 있다.

[0021] 본 개시의 일 실시예에 따른, 음성 합성 방법은, 문자열을 입력받는 단계; 화자 정보를 입력 받은 후 상기 화자 정보를 통해 화자 잠재 벡터를 생성하는 단계; 상기 문자열을 문자 잠재 벡터로 인코딩하는 단계; 상기 문자 잠재 벡터 및 상기 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성하는 단계; 및 상기 문자 잠재 벡터, 상기 화자 잠재 벡터, 및 상기 운율 잠재 벡터를 이용하여 합성 음성을 생성하는 단계를 포함할 수 있다.

[0022] 본 개시의 일 실시예에 따른, 음성 합성 방법은, 참조 음성을 입력받는 단계; 상기 참조 음성을 참조 음성 운율 잠재 벡터로 인코딩하는 단계; 및 상기 합성 음성 및 상기 참조 음성 운율 잠재 벡터에 기초하여 상기 합성 음성이 비-합성 음성인지 여부를 판별하는 단계를 더 포함할 수 있다.

[0023] 그리고, 상기 합성 음성이 비-합성 음성인지 여부를 판별하는 단계는 노이즈 제거 확산 확률 모델을 통해 수행될 수 있다.

[0024] 상기 과제의 해결 수단은, 본 개시의 특징을 모두 열거한 것은 아니다. 본 개시의 다양한 특징과 그에 따른 장점과 효과는 아래의 구체적인 실시예를 참조하여 보다 상세하게 이해될 수 있을 것이다.

발명의 효과

[0025] 본 개시의 다양한 실시예에 의해, 확산 기반 잠재 운율 생성 및 운율 조건부 적대적 훈련을 통한 표현적 음성 합성을 수행하는 방법 및 장치가 제공될 수 있다.

[0026] 본 개시의 다양한 실시예에 의해, 운율을 단어 단위 벡터로 모델링하고, 이를 노이즈 제거 확산 확률 모델을 통해 생성하는 방법 및 표현력 있는 음성 합성을 위한 운율 조건부 적대적 훈련 방법 및 장치가 제공될 수 있다.

[0027] 본 개시의 다양한 실시예에 의해, 참조 음성 없이 화자 정보와 문자열을 이용하여 운율 잠재 벡터를 생성함으로써 표현적인 음성을 합성할 수 있다.

[0028] 본 개시의 다양한 실시예에 의해, 노이즈 제거 확산 확률 모델은 자기 회귀적 방법보다 빠르고 안정적으로 운율 잠재 벡터를 생성하고, 운율 조건부 적대적 훈련은 생성된 음성이 더 표현성 있도록 생성기 모델이 학습되도록 유도될 수 있다.

[0029] 본 개시의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로

부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0030] 도 1은 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하는 장치의 구성을 설명하기 위한 블록도이다.

도 2는 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하기 위한 학습 방법을 설명하기 위한 도면이다.

도 3은 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하기 위한 추론 방법을 설명하기 위한 도면이다.

도 4는 본 개시의 일 실시예에 따른, 운율 잠재 벡터 예측부의 학습 동작을 설명하기 위한 도면이다.

도 5는 본 개시의 일 실시예에 따른, 운율 잠재 벡터 예측부의 추론 동작을 설명하기 위한 도면이다.

도 6은 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하는 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 개시의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시의 개시가 완전하도록 하고, 본 개시가 속하는 기술 분야의 통상의 기술자에게 본 개시의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 개시는 청구항의 범주에 의해 정의될 뿐이다.

[0032] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 개시를 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다.

[0033] 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 개시의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0034] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 개시가 속하는 기술 분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0035] 공간적으로 상대적인 용어인 "아래(below)", "아래(beneath)", "하부(lower)", "위(above)", "상부(upper)" 등은 도면에 도시되어 있는 바와 같이 하나의 구성요소와 다른 구성요소과의 상관관계를 용이하게 기술하기 위해 사용될 수 있다. 공간적으로 상대적인 용어는 도면에 도시되어 있는 방향에 더하여 사용시 또는 동작 시 구성요소들의 서로 다른 방향을 포함하는 용어로 이해되어야 한다.

[0036] 예를 들어, 도면에 도시되어 있는 구성요소를 뒤집을 경우, 다른 구성요소의 "아래(below)"또는 "아래(beneath)"로 기술된 구성요소는 다른 구성요소의 "위(above)"에 놓여질 수 있다. 따라서, 예시적인 용어인 "아래"는 아래와 위의 방향을 모두 포함할 수 있다. 구성요소는 다른 방향으로도 배향될 수 있으며, 이에 따라 공간적으로 상대적인 용어들은 배향에 따라 해석될 수 있다.

[0037] 이하에서는 도면을 참조하여 확산 기반 잠재 운율 생성 및 운율 조건부 적대적 훈련을 통한 표현적 음성 합성을 수행하는 방법을 설명하도록 한다.

[0038] 도 1은 본 개시의 일 실시예에 따른, 도 1은 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하는 장치의 구성을 설명하기 위한 블록도이다.

[0039] 도 1에서 설명하는 장치는 스마트 폰, 태블릿 PC, 웨어러블 장치, 노트북 등을 포함할 수 있으나 이에 제한되는 것은 아니다.

[0040] 도 1에 도시된 바와 같이, 장치(100)는 메모리(110), 통신 모듈(120), 마이크(130) 및 프로세서(140)를 포함할

수 있다. 다만, 이에 국한되는 것은 아니며, 장치(100)는 필요한 동작에 따라 당업자 관점에서 자명한 범위 내에서 소프트웨어 및 하드웨어 구성이 수정/추가/생략될 수 있다.

[0041] 메모리(110)는 본 장치(100)의 다양한 기능을 지원하는 데이터와, 제어부의 동작을 위한 프로그램을 저장할 수 있고, 입/출력되는 데이터들을 저장할 있고, 본 장치에서 구동되는 다수의 응용 프로그램(application program 또는 애플리케이션(application)), 본 장치의 동작을 위한 데이터들, 명령어들을 저장할 수 있다. 이러한 응용 프로그램 중 적어도 일부는, 무선 통신을 통해 외부 장치로부터 다운로드 될 수 있다.

[0042] 이러한, 메모리(110)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), SSD 타입(Solid State Disk type), SDD 타입(Silicon Disk Drive type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(random access memory; RAM), SRAM(static random access memory), 롬(read-only memory; ROM), EEPROM(electrically erasable programmable read-only memory), PROM(programmable read-only memory), 자기 메모리, 자기 디스크 및 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 또한, 메모리(110)는 본 장치와는 분리되어 있으나, 유선 또는 무선으로 연결된 데이터베이스가 될 수도 있다.

[0043] 통신 모듈(120)은 외부 장치와 통신을 가능하게 하는 하나 이상의 구성 요소를 포함할 수 있다. 예를 들어, 통신 모듈(120)은 무선통신 모듈, 유선통신 모듈 또는 위치정보 모듈 중 적어도 하나를 포함할 수 있다.

[0044] 여기서, 무선 통신 모듈은 와이파이(Wifi) 모듈, 와이브로(Wireless broadband) 모듈 외에도, GSM(global System for Mobile Communication), CDMA(Code Division Multiple Access), WCDMA(Wideband Code Division Multiple Access), UMTS(universal mobile telecommunications system), TDMA(Time Division Multiple Access), LTE(Long Term Evolution), 4G, 5G, 6G 등 다양한 무선 통신 방식을 지원하는 무선 통신 모듈을 포함할 수 있다.

[0045] 마이크(130)는 사용자(즉, 발화자)의 음성을 입력받을 수 있다. 도 1에서는 마이크(130)가 장치(100)의 일 구성요소로 구현된 경우를 도시하고 있으나 이에 제한되는 것은 아니다. 마이크(130)는 장치(100)의 외부에 구성될 수 있으며, 장치(100)는 마이크(130)로부터 입력된 발화자의 음성을 통신 모듈(120)을 통해 입력받을 수 있다.

[0046] 프로세서(140)는 본 장치(100) 내의 구성요소들의 동작을 제어하기 위한 알고리즘 또는 알고리즘을 재현한 프로그램에 대한 데이터를 저장하는 메모리, 및 메모리에 저장된 데이터를 이용하여 전술한 동작을 수행하는 적어도 하나의 프로세서(미도시)로 구현될 수 있다. 이때, 메모리와 프로세서는 각각 별개의 칩으로 구현될 수 있다. 또는, 메모리와 프로세서는 단일 칩으로 구현될 수도 있다.

[0047] 즉, 프로세서(140)는 본 장치(100) 내의 구성 요소들을 제어함으로써 장치의 전반적인 동작 및 기능을 제어할 수 있다. 프로세서는 이하의 도면에서 설명되는 본 개시에 따른 다양한 실시 예들을 본 장치 상에서 구현하기 위하여, 위에서 살펴본 구성요소들을 중 어느 하나 또는 복수를 조합하여 제어할 수 있다.

[0048] 프로세서(140)는 문자 입력부(140-1), 문자 인코더부(140-2), 화자 입력부(140-3), 음성 입력부(140-4), 운율 잠재 벡터 인코더(140-5), 운율 잠재 벡터 예측부(140-6), 노이즈 제거부(140-7), 음성 합성부(140-8) 및 운율 조건부 판독부(140-9)를 포함할 수 있다.

[0049] 즉, 프로세서(140)는 문자 입력부(140-1), 문자 인코더부(140-2), 화자 입력부(140-3), 음성 입력부(140-4), 운율 잠재 벡터 인코더(140-5), 운율 잠재 벡터 예측부(140-6), 노이즈 제거부(140-7), 음성 합성부(140-8) 및 운율 조건부 판독부(140-9)를 제어할 수 있다.

[0050] 구체적으로, 문자 입력부(140-1)는 합성할 문자열을 입력받을 수 있다.

[0051] 문자 인코더부(140-2)는 입력된 문자(즉, 입력 문자열)를 문자 잠재 벡터로 인코딩할 수 있다.

[0052] 화자 입력부(140-3)는 화자 정보를 입력 받은 후 이를 이용하여 화자 잠재 벡터를 생성할 수 있다.

[0053] 음성 입력부(140-4)는 운율 잠재 벡터를 모델링할 음성을 입력받을 수 있다.

[0054] 운율 잠재 벡터 인코더부(140-5)는 입력 받은 음성을 운율 잠재 벡터로 인코딩할 수 있다.

[0055] 노이즈 제거부(140-7)는 문자 잠재 벡터와 화자 잠재 벡터, 시간 단계, 노이즈를 입력 받아 시간 단계에 맞는 노이즈를 제거할 수 있다. 즉, 노이즈 제거부(140-7)는 전진 과정에서 학습 운율 잠재 벡터에 대해 노이즈를 부가할 수 있다.

[0056] 음성 합성부(140-8)는 문자 잠재 벡터, 화자 잠재 벡터 및 운율 잠재 벡터를 입력받아 음성을 합성할 수 있다.

[0057] 운율 조건부 판독부(140-9)는 음성 및 운율 잠재 벡터를 입력 받아 진짜 데이터인지 생성된 데이터인지 여부를 판별할 수 있다.

[0058] 도 2는 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하기 위한 학습 방법을 설명하기 위한 도면이다. 즉, 도 2는 음성 합성부(140-8) 및 운율 잠재 벡터 인코더(또는, 운율 벡터 인코더부)(140-5)의 학습 과정을 설명하기 위한 도면이다.

[0059] 음성 합성부(140-8)은 음성으로 합성할 문자열, 화자 정보, 운율 모델링을 위한 음성에 기초하여 음성을 생성할 수 있다.

[0060] 이 때, 문자열은 문자 입력부(140-1)를 통해 입력받을 수 있다. 화자 정보는 화자 입력부(140-3)를 통해 입력받을 수 있다. 운율 모델링을 위한 음성은 음성 입력부(140-4)를 통해 입력받을 수 있다.

[0061] 운율 조건부 판별부(140-9)는 음성 및 운율 잠재 벡터를 입력 받아 운율 조건부 적대적 훈련을 수행할 수 있다.

[0062] 도 3은 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하기 위한 추론 방법을 설명하기 위한 도면이다.

[0063] 음성 합성부(140-8)는 음성으로 합성할 문자열 및 화자 정보를 입력 받아 음성을 합성/생성할 수 있다. 여기서, 운율 잠재 벡터 예측부(또는, 운율 벡터 예측부)(140-6)는 문자 잠재 벡터 및 화자 잠재 벡터를 입력 받아 운율 잠재 벡터를 생성할 수 있다. 여기서, 도 2를 참조하여 설명한 운율 잠재 벡터 인코더부는 사용되지 않을 수 있다.

[0064] 즉, 도 3에서는 운율 잠재 벡터 인코더부 없이 음성 합성부(140-8)가 문자 잠재 벡터, 화자 잠재 벡터 및 운율 잠재 벡터를 통해 음성을 합성할 수 있다.

[0065] 도 4는 본 개시의 일 실시예에 따른, 운율 잠재 벡터 예측부의 학습 동작을 설명하기 위한 도면이다. 즉, 도 4는 운율 잠재 벡터 예측부(140-6)의 학습을 나타내는 흐름도이다.

[0066] 운율 잠재 벡터 예측부(140-6)는 노이즈 제거 확산 확률 모델을 의미하며, 전진 과정 및 후진 과정이 적용될 수 있다.

[0067] 일 예로, 전진 과정은 운율 잠재 벡터에 노이즈를 더하는 과정을 의미하며, 후진 과정은 노이즈가 더해진 운율 잠재 벡터에서 노이즈를 제거하는 과정을 의미한다.

[0068] 후진 과정을 위해서 노이즈 제거부(140-7)가 이용될 수 있다. 노이즈 제거부(140-7)는 문자 잠재 벡터, 화자 잠재 벡터, 및 시간 단계를 입력 받아 전진 과정으로 만들어진 데이터 상에서 노이즈를 제거할 수 있다.

[0069] 여기서, 문자 잠재 벡터는 입력부(140-1)에 의해 입력된 문자를 인코딩함으로써 생성될 수 있다. 화자 잠재 벡터는 화자 입력부(140-3)에 의해 생성될 수 있다. 시간 단계는 시간 단계 입력부(140-11)에 의해 생성될 수 있다.

[0070] 도 5는 본 개시의 일 실시예에 따른, 운율 잠재 벡터 예측부의 추론 동작을 설명하기 위한 도면이다. 즉, 도 5는 운율 잠재 벡터 예측부의 추론을 나타내는 흐름도이다.

[0071] 노이즈 제거부(140-7)는 문자 잠재 벡터, 화자 잠재 벡터, 시간 단계 노이즈를 입력 받아 점진적으로 노이즈를 제거함으로써 운율 잠재 벡터를 생성할 수 있다.

[0072] 도 6은 본 개시의 일 실시예에 따른, 표현적 음성 합성을 수행하는 방법을 설명하기 위한 순서도이다.

[0073] 장치는 문자열을 문자 입력부를 통해 입력받을 수 있다(S610).

[0074] 그리고, 장치는 화자 입력부를 통해 화자 정보를 입력 받은 후 화자 정보를 통해 화자 잠재 벡터를 획득할 수 있다(S620).

[0075] 도 6에서는 S610 단계 이후 S620 단계가 진행되는 경우를 도시하고 있으나 이에 제한되는 것은 아니다. S610 단계 및 S620 단계는 동시에 진행될 수도 있으며 S620 단계가 먼저 진행될 수도 있다.

[0076] 장치는 문자 인코더부를 통해 문자열을 문자 잠재 벡터로 인코딩할 수 있다(S630).

[0077] 장치는 운율 잠재 벡터 예측부를 통해 문자 잠재 벡터 및 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성할 수 있다(S640).

[0078] 여기서, 운율 잠재 벡터 예측부는 노이즈 제거 확산 확률 모델을 포함할 수 있다. 노이즈 제거 확산 확률 모델은 전진 과정(forward process)를 통해 학습 운율 잠재 벡터에 대해 노이즈를 부가하고, 후진 과정(reverse process)를 통해 노이즈가 부가된 학습 운율 잠재 벡터에 대해 노이즈를 제거하도록 학습될 수 있다.

[0079] 이 때, 장치는 노이즈 입력부를 통해 운율 잠재 벡터에 대해 노이즈를 부가하고, 노이즈 제거부를 통해 노이즈가 부가된 운율 잠재 벡터에 대해 노이즈를 제거할 수 있다.

[0080] 여기서, 노이즈 입력부 또는 노이즈 제거부 중의 적어도 하나는, 합성곱 신경망(convolutional neural network, CNN), 순환 신경망(recurrent neural network, RNN), 장-단기 기억(long short term memory, LSTM), GRU(gated-recurrent unit) 또는 트랜스포머(transformer) 중의 적어도 하나를 포함할 수 있다.

[0081] 뿐만 아니라, 운율 잠재 벡터 인코더 및/또는 운율 잠재 벡터 예측부 등은 합성곱 신경망(convolutional neural network, CNN), 순환 신경망(recurrent neural network, RNN), 장-단기 기억(long short term memory, LSTM), GRU(gated-recurrent unit) 또는 트랜스포머(transformer) 중의 적어도 하나를 포함할 수 있다.

[0082] 장치는 문자 잠재 벡터, 화자 잠재 벡터, 및 운율 잠재 벡터를 이용하여 합성 음성을 생성할 수 있다(S650).

[0083] 추가적으로, 장치는 음성 입력부를 통해 참조 음성을 더 입력받을 수 있다. 장치는 참조 음성-운율 벡터 인코더부를 통해 참조 음성을 참조 음성 운율 잠재 벡터로 인코딩할 수 있다. 장치는 운율 조건부 판별부를 통해 합성 음성 및 참조 음성 운율 잠재 벡터에 기초하여 합성 음성이 비-합성 음성인지 여부를 판별할 수 있다.

[0084] 여기서, 문자 인코더부 및 합성 음성부는 운율 조건 판별부에 의해 적대적 조건 훈련 방식을 통해 합성 음성이 비-합성 음성으로 판별되도록 학습될 수 있다.

[0085] 이상, 본 개시에서 설명되는 실시예와 첨부된 도면은 본 개시에 포함되는 기술적 사상의 일부를 예시적으로 설명하는 것에 불과하다. 따라서, 본 명세서에 개시된 실시예들은 본 개시의 기술적 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이므로, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것이 아님은 자명하다.

[0086] 본 개시의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시 예는 모두 본 개시의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0087] 100 : 장치

110 : 메모리

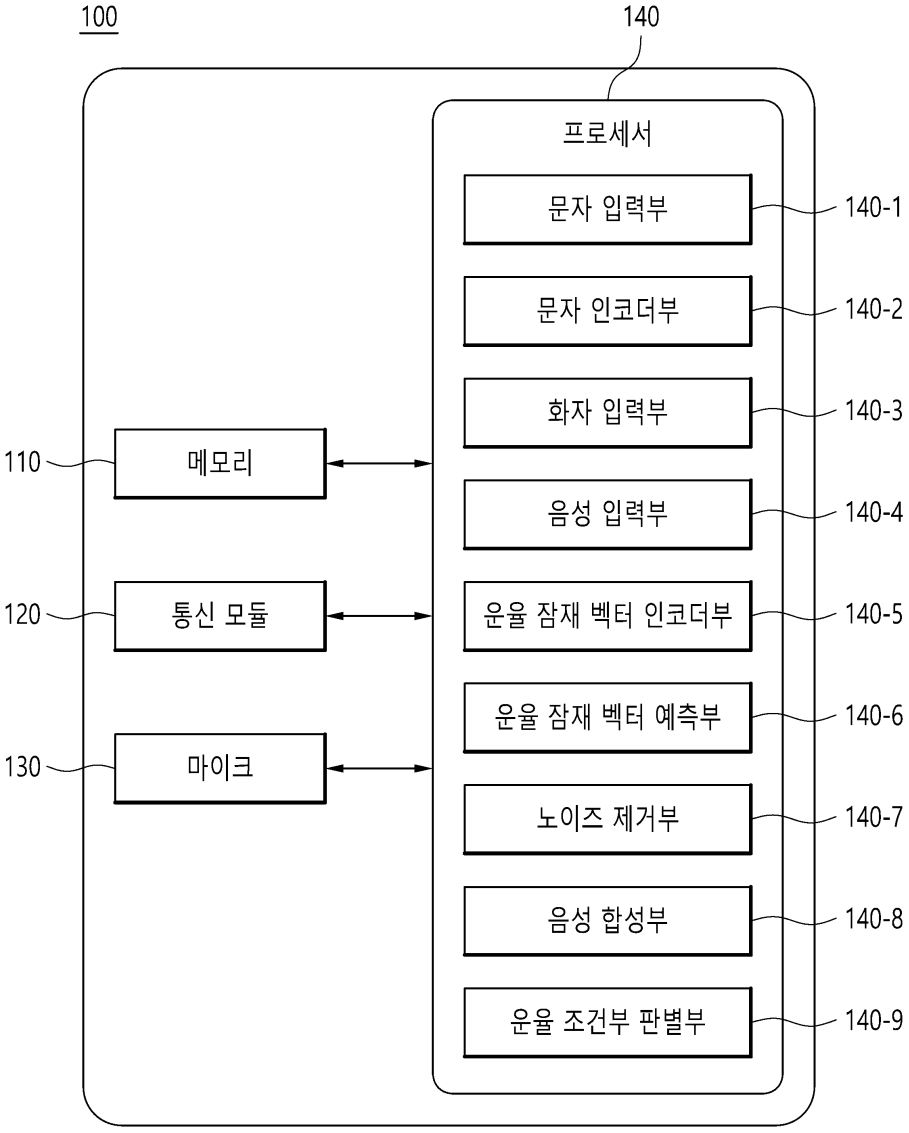
120 : 통신 모듈

130 : 마이크

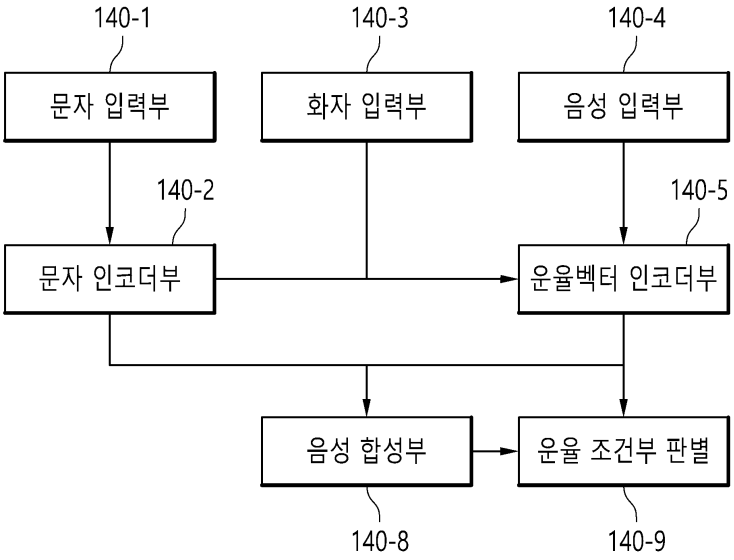
140 : 프로세서

도면

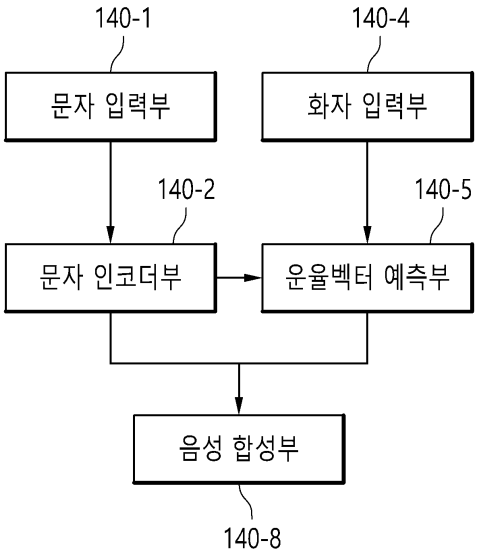
도면1


100
140
프로세서
문자 입력부
140-1
문자 인코더부
140-2
화자 입력부
140-3
음성 입력부
140-4
운율 잠재 벡터 인코더부
140-5
운율 잠재 벡터 예측부
140-6
노이즈 제거부
140-7
음성 합성부
140-8
운율 조건부 판별부
140-9
110
메모리
120
통신 모듈
130
마이크

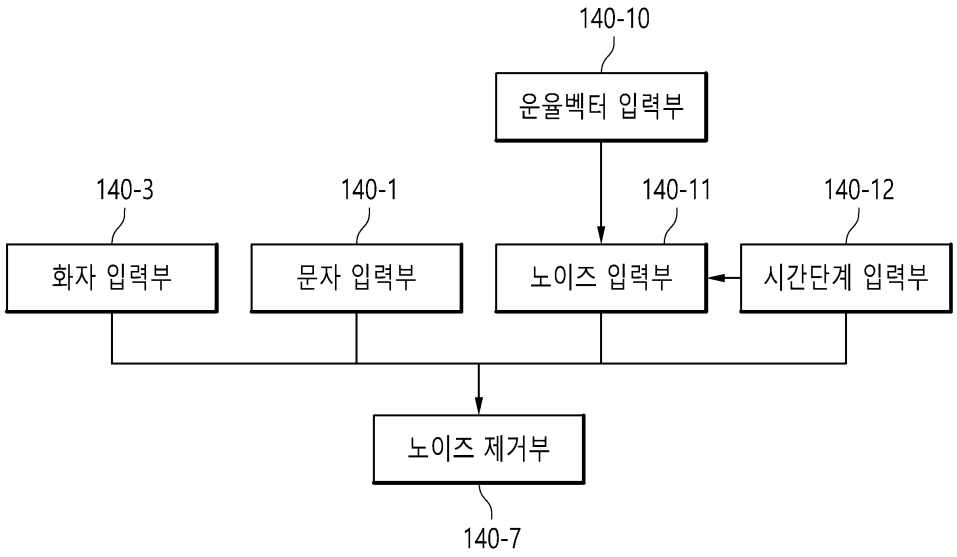
도면2


140-1
140-3
140-4
문자 입력부
화자 입력부
음성 입력부
140-2
140-5
문자 인코더부
운율벡터 인코더부
음성 합성부
운율 조건부 판별
140-8
140-9

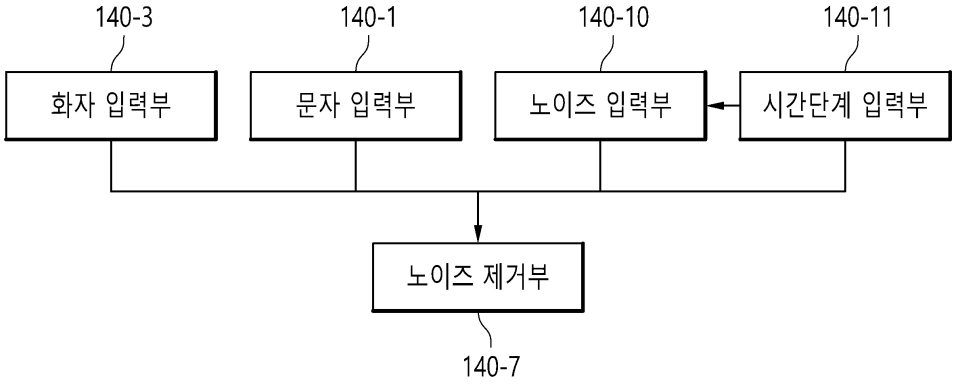
도면3


140-1
140-4
문자 입력부
화자 입력부
140-2
140-5
문자 인코더부
운율벡터 예측부
음성 합성부
140-8

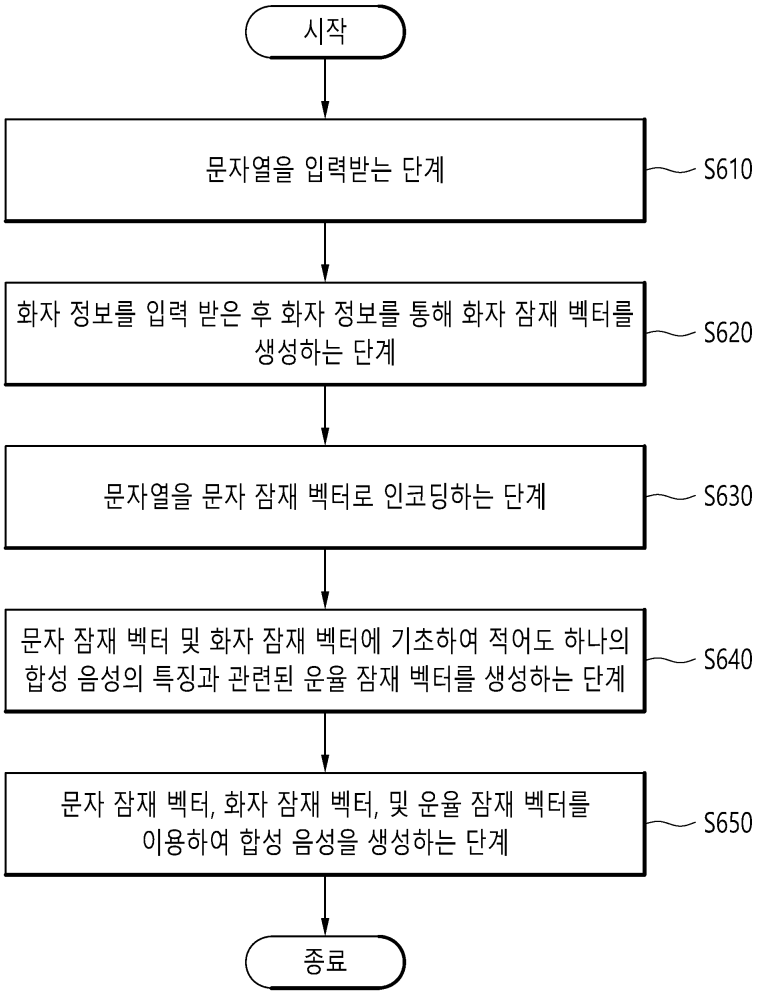
도면4


140-10
운율벡터 입력부
140-3
140-1
140-11
140-12
화자 입력부
문자 입력부
노이즈 입력부
시간단계 입력부
노이즈 제거부
140-7

도면5


140-3
140-1
140-10
140-11
화자 입력부
문자 입력부
노이즈 입력부
시간단계 입력부
노이즈 제거부
140-7

도면6


시작
문자열을 입력받는 단계
S610
화자 정보를 입력 받은 후 화자 정보를 통해 화자 잠재 벡터를 생성하는 단계
S620
문자열을 문자 잠재 벡터로 인코딩하는 단계
S630
문자 잠재 벡터 및 화자 잠재 벡터에 기초하여 적어도 하나의 합성 음성의 특징과 관련된 운율 잠재 벡터를 생성하는 단계
S640
문자 잠재 벡터, 화자 잠재 벡터, 및 운율 잠재 벡터를 이용하여 합성 음성을 생성하는 단계
S650
종료