(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년02월16일
(11) 등록번호 10-2637341
(24) 등록일자 2024년02월13일

(51) 국제특허분류(Int. Cl.)
G10L 13/08 (2006.01) *G10L 15/08* (2006.01)
G10L 15/16 (2006.01)
(52) CPC특허분류
G10L 13/08 (2013.01)
G10L 15/08 (2013.01)
(21) 출원번호 10-2019-0127701
(22) 출원일자 2019년10월15일
심사청구일자 2022년08월31일
(65) 공개번호 10-2021-0044484
(43) 공개일자 2021년04월23일
(56) 선행기술조사문헌
KR1020120048823 A*
KR1020190094314 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
이장수
서울특별시 서초구 잠원로8길 20, 330동 105호 (잠원동, 신반포19차아파트)
이호식
경기도 성남시 분당구 내정로 152 ,127동1601호(수내동,파크타운)
전재훈
경기도 수원시 영통구 매탄로126번길 66, 205동 1306호 (매탄동, 주공그린빌)
(74) 대리인
특허법인 무한

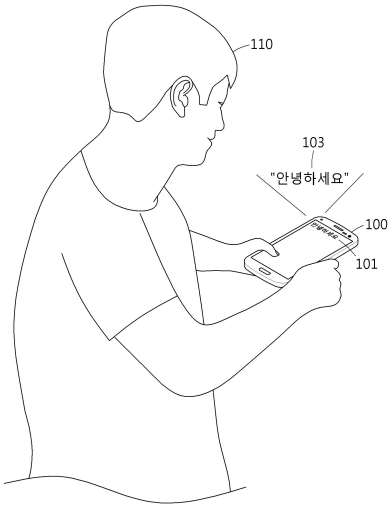
전체 청구항 수 : 총 19 항 심사관 : 정성윤

(54) 발명의 명칭 **음성 생성 방법 및 장치**

(57) 요 약

음성 생성 방법 및 장치가 개시된다. 음성 생성 방법은, 프로세서에 의해, 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하는 단계, 프로세서에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하는 단계, 프로세서에 의해, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계 및 프로세서에 의해, 상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 15/16 (2013.01)
G10L 2015/081 (2013.01)

명 세 서

청구범위

청구항 1

프로세서에 의해, 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하는 단계;

상기 프로세서에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하는 단계;

상기 프로세서에 의해, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계; 및

상기 프로세서에 의해, 상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력하는 단계를 포함하고,

상기 제2 후보 음성 유닛을 생성하는 단계는,

상기 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계를 포함하는,

음성 생성 방법.

청구항 2

제1항에 있어서,

상기 제2 후보 음성 유닛을 생성하는 단계는,

상기 음성 유닛 생성기에 의해, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제3 후보 음성 유닛으로부터 스타일(style) 특징을 추출하는 단계;

상기 음성 유닛 생성기에 의해, 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제4 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출하는 단계; 및

상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징, 상기 스타일 특징 및 상기 컨텐츠 특징을 기초로 상기 제2 후보 음성 유닛을 생성하는 단계

를 포함하는, 음성 생성 방법.

청구항 3

제1항에 있어서,

상기 제2 후보 음성 유닛을 생성하는 단계는,

상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 상기 제2 후보 음성 유닛을 생성하는 단계를 포함하고,

상기 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 포함하는, 음성 생성 방법.

청구항 4

프로세서에 의해, 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하는 단계;

상기 프로세서에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하는 단계;

상기 프로세서에 의해, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계;

상기 프로세서에 의해, 상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력하는 단계; 및

제1 후보 음성 시퀀스를 구성하는 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 복수의 후보 음성 유닛을 생성하여 메모리에 저장하는 단계

를 포함하는, 음성 생성 방법.

청구항 5

제1항에 있어서,

상기 제2 후보 음성 유닛을 생성하는 단계는,

상기 입력 텍스트에 대응하는 각각의 음성 단위 마다 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는,

음성 생성 방법.

청구항 6

삭제

청구항 7

프로세서에 의해, 학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하는 단계;

프로세서에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 함수를 이용한 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하는 단계;

프로세서에 의해, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계;

프로세서에 의해, 상기 제1 후보 음성 유닛 및 상기 제2 후보 음성 유닛을 기초로 상기 제2 후보 음성 유닛에 대응하는 손실값을 계산하는 단계; 및

프로세서에 의해, 상기 손실값을 기초로 상기 음성 유닛 생성기의 파라미터를 갱신하는 단계를 포함하고,

상기 제2 후보 음성 유닛을 생성하는 단계는,

상기 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계를 포함하는,

음성 유닛 생성기 학습 방법.

청구항 8

제7항에 있어서,

상기 손실값을 계산하는 단계는,

상기 코스트 함수가 미분 가능한 경우, 상기 코스트 함수를 이용하여 상기 손실값을 계산하는 단계를 포함하는, 음성 유닛 생성기 학습 방법.

청구항 9

제7항에 있어서,

상기 갱신하는 단계는,

복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제2 후보 음성 유닛으로부터 스타일(style) 특징을 스타일 추출기의 파라미터 및 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제3 후보 음성 유닛으로부터 컨텐츠(content) 특징을 특징 추출기를 포함하는 상기 음성 유닛 생성기의 파라미터를 갱신하는 단계

를 포함하는, 음성 유닛 생성기 학습 방법.

청구항 10

제7항에 있어서,

상기 갱신하는 단계는,

상기 음성 유닛 생성기에 포함된 생성자 (Generator) 및 감별자 (Discriminator)의 파라미터를 갱신하는 단계를 포함하는, 음성 유닛 생성기 학습 방법.

청구항 11

컴퓨팅 하드웨어가 제1항 내지 제5항 및 제7항 내지 제9항 중 어느 하나의 항의 방법을 실행하도록 하는 인스트럭션들을 저장하는 비일시적인(non-transitory) 컴퓨터 판독 가능한 저장 매체.

청구항 12

적어도 하나의 프로세서; 및

음성 유닛 생성기를 저장하는 메모리를 포함하고,

상기 프로세서는,

입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하고,

상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하고,

상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하고,

상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력하고,

상기 프로세서는,

상기 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는,

음성 생성 장치.

청구항 13

제12항에 있어서,

상기 프로세서는,

상기 음성 유닛 생성기에 의해, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제3 후보 음성 유닛으로부터 스타일(style) 특징을 추출하고,

상기 음성 유닛 생성기에 의해, 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제4 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출하고,

상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징, 상기 스타일 특징 및 상기 컨텐츠 특징을 기초로 상기 제2 후보 음성 유닛을 생성하는,

음성 생성 장치.

청구항 14

제12항에 있어서,

상기 프로세서는,

상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 상기 제2 후보 음성 유닛을 생성하고,

상기 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 포함하는, 음성 생성 장치.

청구항 15

적어도 하나의 프로세서; 및

음성 유닛 생성기를 저장하는 메모리를 포함하고,

상기 프로세서는,

입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하고,

상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하고,

상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하고,

상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력하고,

상기 프로세서는,

제1 후보 음성 시퀀스를 구성하는 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 복수의 후보 음성 유닛을 생성하여 상기 메모리에 저장하는,

음성 생성 장치.

청구항 16

제12항에 있어서,

상기 프로세서는,

상기 입력 텍스트에 대응하는 각각의 음성 단위 마다 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는,

음성 생성 장치.

청구항 17

삭제

청구항 18

적어도 하나의 프로세서; 및

음성 유닛 생성기를 저장하는 메모리를 포함하고,

상기 프로세서는,

학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하고,

상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 함수를 이용한 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하고,

상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하고,

상기 제1 후보 음성 유닛 및 상기 제2 후보 음성 유닛을 기초로 상기 제2 후보 음성 유닛에 대응하는 손실값을 계산하고,

상기 손실값을 기초로 상기 음성 유닛 생성기의 파라미터를 갱신하고,

상기 프로세서는,

상기 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는,

음성 유닛 생성기 학습 장치.

청구항 19

제18항에 있어서,

상기 프로세서는,

상기 코스트 함수가 미분 가능한 경우, 상기 코스트 함수를 이용하여 상기 손실값을 계산하는, 음성 유닛 생성기 학습 장치.

청구항 20

제18항에 있어서,

상기 프로세서는,

복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제2 후보 음성 유닛으로부터 스타일(style) 특징을 스타일 추출기의 파라미터 및 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제3 후보 음성 유닛으로부터 컨텐츠(content) 특징을 특징 추출기를 포함하는 상기 음성 유닛 생성기의 파라미터를 갱신

하는,

음성 유닛 생성기 학습 장치.

청구항 21

제18항에 있어서,

상기 프로세서는,

상기 음성 유닛 생성기에 포함된 생성자 (Generator) 및 감별자 (Discriminator)의 파라미터를 갱신하는, 음성 유닛 생성기 학습 장치.

발명의 설명

기 술 분 야

[0001] 음성 합성 기술에 관한 것으로, 입력 텍스트로부터 음성 유닛을 생성하여 자연스러운 출력 음성을 출력하는 기술에 관한 것이다.

배 경 기 술

[0002] 음성 합성(speech synthesis)은 말소리의 음파를 기계가 자동으로 만들어 내는 기술로, 간단히 말하면 모델로 선정된 한 사람의 말소리를 녹음하여 일정한 음성 단위로 분할한 다음, 부호를 붙여 합성기에 입력하였다가 지시에 따라 필요한 음성 단위만을 다시 합쳐 말소리를 인위로 만들어내는 기술이다. 음성 합성은 TTS(Text To Speech)로 지칭될 수 있다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0003] 일 실시예에 따른 음성 생성 방법은, 프로세서에 의해, 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하는 단계; 프로세서에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하는 단계; 프로세서에 의해, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계; 및 프로세서에 의해, 상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력하는 단계를 포함한다.

[0004] 상기 제2 후보 음성 유닛을 생성하는 단계는, 상기 음성 유닛 생성기에 의해, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제3 후보 음성 유닛으로부터 스타일(style) 특징을 추출하는 단계; 상기 음성 유닛 생성기에 의해, 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제4 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출하는 단계; 및 상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징, 상기 스타일 특징 및 상기 컨텐츠 특징을 기초로 상기 제2 후보 음성 유닛을 생성하는 단계를 포함할 수 있다.

[0005] 상기 제2 후보 음성 유닛을 생성하는 단계는, 상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 상기 제2 후보 음성 유닛을 생성하는 단계를 포함하고, 상기 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 포함할 수 있다.

[0006] 상기 방법은, 상기 제1 후보 음성 시퀀스를 구성하는 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 복수의 후보 음성 유닛을 생성하여 메모리에 저장하는 단계를 더 포함할 수 있다.

[0007] 상기 제2 후보 음성 유닛을 생성하는 단계는, 상기 입력 텍스트에 대응하는 각각의 음성 단위 마다 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다.

[0008] 상기 제2 후보 음성 유닛을 생성하는 단계는, 상기 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계를 포함할 수 있다.

[0009] 일 실시예에 따른 학습 방법은, 프로세서에 의해, 학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하는 단계; 프로세서에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하는 단계; 프로세서에 의해, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하는 단계; 프로세서에 의해, 상기 제1 후보 음성 유닛 및 상기 제2 후보 음성 유닛을 기초로 상기 제2 후보 음성 유닛에 대응하는 손실값을 계산하는 단계; 및 프로세서에 의해, 상기 손실값을 기초로 상기 음성 유닛 생성기의 파라미터를 갱신하는 단계를 포함할 수 있다.

[0010] 상기 손실값을 계산하는 단계는, 상기 코스트 함수가 미분 가능한 경우, 상기 코스트 함수를 이용하여 상기 손실값을 계산하는 단계를 포함할 수 있다.

[0011] 상기 갱신하는 단계는, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제2 후보 음성 유닛으로부터 스타일(style) 특징을 스타일 추출기의 파라미터 및 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제3 후보 음성 유닛으로부터 컨텐츠(content) 특징을 특징 추출기를 포함하는 상기 음성 유닛 생성기의 파라미터를 갱신하는 단계를 포함할 수 있다.

[0012] 상기 갱신하는 단계는, 상기 음성 유닛 생성기에 포함된 생성자 (Generator) 및 감별자 (Discriminator)의 파라미터를 갱신하는 단계를 포함할 수 있다.

[0013] 일 실시예에 따른 비일시적인(non-transitory) 컴퓨터 판독 가능한 저장 매체는 컴퓨팅 하드웨어가 제1항 내지 제9항 중 어느 하나의 항의 방법을 실행하도록 하는 인스트럭션들을 저장할 수 있다.

[0014] 일 실시예에 따른 음성 생성 장치는, 적어도 하나의 프로세서; 및 음성 유닛 생성기를 저장하는 메모리를 포함하고, 상기 프로세서는, 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하고, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하고, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하고, 상기 제2 후보 음성 유닛을 상기 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력할 수 있다.

[0015] 상기 프로세서는, 상기 음성 유닛 생성기에 의해, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제3 후보 음성 유닛으로부터 스타일(style) 특징을 추출하고, 상기 음성 유닛 생성기에 의해, 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제4 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출하고, 상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징, 상기 스타일 특징 및 상기 컨텐츠 특징을 기초로 상기 제2 후보 음성 유닛을 생성할 수 있다.

[0016] 상기 프로세서는, 상기 음성 유닛 생성기에 의해, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 상기 제2 후보 음성 유닛을 생성하고, 상기 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 포함할 수 있다.

[0017] 상기 프로세서는, 상기 제1 후보 음성 시퀀스를 구성하는 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 복수의 후보 음성 유닛을 생성하여 상기 메모리에 저장할 수 있다.

[0018] 상기 프로세서는, 상기 입력 텍스트에 대응하는 각각의 음성 단위 마다 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다.

[0019] 상기 프로세서는, 상기 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 상기 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다.

[0020] 일 실시예에 따른 학습 장치는, 적어도 하나의 프로세서; 및 음성 유닛 생성기를 저장하는 메모리를 포함하고, 상기 프로세서는, 학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득하고, 상기 링귀스틱 특징 및 상기 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정하고, 상기 링귀스틱 특징 또는 상기 프로소디 특징 및 상기 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성하고, 상기 제1 후보 음성 유닛 및 상기 제2 후보 음성 유닛을 기초로 상기 제2 후보 음성 유닛에 대응하는 손실값을 계산하고, 상기 손실값을 기초로 상기 음성 유닛 생성기의 파라미터를

갱신한다.

[0021] 상기 프로세서는, 상기 코스트 함수가 미분 가능한 경우, 상기 코스트 함수를 이용하여 상기 손실값을 계산할 수 있다.

[0022] 상기 프로세서는, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제2 후보 음성 유닛으로부터 스타일(style) 특징을 스타일 추출기의 파라미터 및 상기 복수의 후보 음성 유닛 중에서 상기 제2 후보 음성 유닛과 상이한 제3 후보 음성 유닛으로부터 컨텐츠(content) 특징을 특징 추출기를 포함하는 상기 음성 유닛 생성기의 파라미터를 갱신할 수 있다.

[0023] 상기 프로세서는, 상기 음성 유닛 생성기에 포함된 생성자 (Generator) 및 감별자 (Discriminator)의 파라미터를 갱신할 수 있다.

도면의 간단한 설명

[0024] 도 1은 일 실시예에 따른 음성 생성 장치에 의해 입력 텍스트로부터 출력 음성이 출력되는 상황을 도시한 도면이다.

도 2는 일 실시예에 따른 음성 생성 방법의 동작을 도시한 순서도이다.

도 3은 일 실시예에 따른 음성 생성 방법의 동작을 도시한 흐름도이다.

도 4는 일 실시예에 따른 음성 생성 방법의 일례를 도시한 흐름도이다.

도 5는 다른 실시예에 따른 음성 생성 방법의 일례를 도시한 흐름도이다.

도 6은 일 실시예에 따른 음성 유닛 생성기의 학습 방법의 동작을 도시한 순서도이다.

도 7은 일 실시예에 따른 음성 생성 장치의 구성을 도시한 도면이다.

도 8은 일 실시예에 따른 음성 유닛 생성기의 학습 장치의 구성을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0025] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 실시될 수 있다. 따라서, 실시예들은 특정한 개시형태로 한정되는 것이 아니며, 본 명세서의 범위는 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0026] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0028] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0029] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0030] 한편, 어떤 실시예가 달리 구현 가능한 경우에 특정 블록 내에 명기된 기능 또는 동작이 순서도와 다르게 수행될 수 있다. 예를 들어, 연속하는 두 블록들이 실제로는 실질적으로 동시에 수행될 수도 있고, 관련된 기능 또는 동작에 따라서는 해당 블록들의 순서가 뒤바뀌어 수행될 수도 있다.

[0031] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한

다.

[0033] 도 1은 일 실시예에 따른 음성 생성 장치에 의해 입력 텍스트로부터 출력 음성이 출력되는 상황을 도시한 도면이다.

[0034] 일 실시예에 따르면, 음성 생성 장치(100)는 입력 텍스트를 입력 받아 입력 텍스트에 대응하는 출력 음성을 출력할 수 있다. 음성 생성 장치(100)는 학습하지 못한 입력 텍스트에 대해서 새로운 음성 유닛을 생성하여 다른 음성 유닛들과 결합할 수 있다. 이처럼, 음성 생성 장치(100)는 학습 과정에서 학습하지 못한 학습 데이터가 입력될 경우에도 적절한 음성 유닛을 생성함으로써 자연스러운 출력 음성을 출력할 수 있다.

[0035] 음성 생성 장치(100)는 음성 합성을 구현하는 장치로서, 스마트폰, DTV, AI Speaker, 전장 또는 로봇 등을 포함할 수 있다. 음성 생성 장치(100)는 음성 합성을 인터페이스로 수행하는 모든 종류의 장치를 포함할 수 있다. 음성 생성 장치(100)는 칩셋 형태의 반도체를 포함할 수 있다.

[0036] 일 실시예에 따른 음성 생성 장치(100)에 적용된 기술은 단지 음성 합성에만 적용되는 것이 아니라, 음성 이외에 다른 이형의 소스(영상, 이미지, 음악 등)를 순차적으로 연결하여 하나의 결과물을 생성하는 경우에 적용될 수 있다. 예를 들어, 일 실시예에 따른 음성 생성 장치(100)에 적용된 기술은 영상이나 이미지의 손상된 부분의 앞 프레임과 뒤 프레임을 고려하여 손상된 부분의 프레임을 생성하고 전체 프레임에 자연스럽게 연결할 수 있다.

[0037] 음성 생성 장치(100)는 음성 합성을 위하여 음성 유닛 생성기를 포함할 수 있다. 음성 유닛 생성기는 뉴럴 네트워크로 구성될 수 있다. 뉴럴 네트워크는 인간의 뉴런의 동작 원리를 모방하여 컴퓨터 상에서 동작하도록 구축된 신경망을 의미한다. 각 뉴런은 입력값에 가중치를 곱하고 편향을 더한 뒤 활성화 함수를 적용하여 출력값을 도출한다. 활성화 함수는 인공 신경망을 통과한 값의 형태를 결정한다. 뉴럴 네트워크는 이러한 뉴런을 복수로 포함하며, 복수의 뉴런들은 입력 레이어, 히든 레이어(hidden layer) 및 출력 레이어를 구성한다. 출력 레이어가 출력한 결과의 오차는 신경망을 따라 역으로 전파하며 각 뉴런들의 파라미터가 조정된다. 이러한 방식을 역전파(backpropagation) 방식이라고 한다.

[0038] 음성 생성 장치(100)는 복수의 음성 유닛을 연결(concatenate)하는 TTS 방식을 사용하여 출력 음성을 생성할 수 있다. 음성 생성 장치(100)는 입력 텍스트의 링귀스틱 특징(linguistic feature)을 추출하고, 프로소디 특징(prosody feature)을 생성할 수 있다. 데이터베이스에는 미리 다양한 링귀스틱 특징 및 프로소디 특징에 대응하는 다수의 음성 유닛(voice unit)이 저장될 수 있다. 음성 생성 장치(100)는 추출된 링귀스틱 특징 및 프로소디 특징에 대응하는 음성 유닛을 데이터베이스에서 선택할 수 있다. 음성 생성 장치(100)는 추출된 링귀스틱 특징 및 프로소디 특징에 가장 적합한 음성 유닛을 선택할 수 있다.

[0039] 입력 텍스트가 복수의 음성 단위에 대응하는 텍스트로 구성된 경우, 음성 생성 장치(100)는 각각의 음성 단위에 대응하여 음성 유닛을 선택할 수 있다. 음성 생성 장치(100)는 음성 유닛을 선택할 때 각 음성 유닛의 코스트(cost)를 계산하고 비터비(Viterbi) 검색을 수행할 수 있다. 음성 생성 장치(100)는 코스트 계산과 비터비 검색을 통해 선택된 복수의 음성 유닛을 연결(concatenate)하여 출력 음성을 생성할 수 있다. 출력 음성은 각 음성 유닛이 자연스럽게 연결된 상태일 수 있다. 여기서, 음성 단위는 포네틱 트랜스크립션(phonetic transcription)의 분절 단위를 지칭할 수 있다.

수학식 1

[0040] $$C = w_t \sum_{n=1}^{N} T(u_n) + w_c \sum_{n=1}^{N} C(u_n, u_{n+1})$$

[0041] 음성 생성 장치(100)는 수학식 1을 이용하여 음성 유닛의 코스트를 계산할 수 있다. 수학식 1에서, 첫 번째 텀(term)은 n 개의 음성 유닛 각각의 코스트의 합을 나타내고, 두 번째 텀은 각 음성 유닛 간의 결합에 대한 코스트의 합을 나타낸다. 음성 생성 장치(100)는 첫 번째 텀과 두 번째 텀의 가중 평균을 통해 C를 계산할 수 있다.

[0042] 하지만, 현실적으로 데이터베이스는 모든 종류의 음성 유닛을 구비할 수 없기 때문에, 주어진 입력 텍스트로부터 추출한 링귀스틱 특징 또는 이에 대응하는 프로소디 특징에 적합한 음성 유닛이 데이터베이스에서 검색되지 않을 수 있다. 음성 생성 장치(100)는 링귀스틱 특징 및 프로소디 특징에 적합한 음성 유닛을 생성할 수 있다.

예를 들어, 음성 생성 장치(100)는 가장 작은 코스트를 가지는 음성 유닛으로부터 스타일 특징(style feature)과 컨텐츠 특징(content featrure)을 추출하고 이들을 합성하여 음성 유닛을 새로 생성할 수 있다. 다른 예로, 음성 생성 장치(100)는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 이용하여 음성 유닛을 생성할 수도 있다. 다만, 음성 유닛의 생성 방법은 이에 국한되지 않으며, 다양한 생성 기법이 적용될 수 있다.

[0043] 프로소디 특징과 유사한 음성 유닛이 데이터베이스에 없는 경우에, 음성 생성 장치(100)는 최소 코스트를 가지는 음성 유닛으로부터 스타일 특징을 추출할 수 있다. 음성 생성 장치(100)는 링귀스틱 특징과 비터비 검색 결과를 기초로 컨텐츠 특징을 추출할 수 있다. 음성 생성 장치(100)는 최소 코스트를 가지는 음성 유닛 이외의 음성 유닛을 이용하여 컨텐츠 특징을 추출할 수 있다. 음성 생성 장치(100)는 컨텐츠 특징에 스타일 특징을 합성하여 새로운 음성 유닛을 생성할 수 있다. 음성 생성 장치(100)는 새로운 음성 유닛을 해당 음성 단위의 후보 음성 유닛의 하나로 설정하고, 비터비 검색을 통해 음성 시퀀스를 획득할 수 있다. 여기서, 스타일 특징은 프로소디 특징을 포함하고, 컨텐츠 특징은 링귀스틱 특징을 포함한다.

[0044] 생성적 적대 신경망은 학습 데이터에 대해 두 개의 뉴럴 네트워크를 경쟁적으로 학습시킴으로써 도출될 수 있다. 학습 과정에 사용되는 두 개의 뉴럴 네트워크는 생성자(Generator) 및 감별자(discriminator)로 지칭될 수 있다. 다만, 이러한 명칭은 식별을 위한 것이며 동일한 기능을 수행한다면 다른 명칭으로 지칭될 수 있다. 생성자는 주어진 학습 데이터로부터 진짜 같은 데이터를 생성할 수 있다. 감별자는 생성자가 생성한 데이터가 진짜인지 아닌지를 판별할 수 있다. 판별 결과를 기초로 역전파 기법을 통해 생성자와 감별자의 기능이 보다 고도화되도록 생성자와 감별자의 파라미터는 갱신될 수 있다.

[0045] 도 1을 참조하면, 음성 생성 장치(100)는 사용자(110)로부터 "안녕하세요"라는 입력 텍스트(101)를 입력 받을 수 있다. 음성 생성 장치(100)는 "안녕하세요"를 분석하여 링귀스틱 특징을 "ㅇ ㅏ ㄴ ㄴ ㅕ ㅇ …"으로 추출할 수 있다. 음성 생성 장치(100)는 링귀스틱 특징에 대응하는 프로소디 특징을 추출할 수 있다. 프로소디 특징은 어쿠스틱 특징(acoustic feature)를 포함할 수 있다. 프로소디 특징은 피치(pitch), 파워(power), 듀레이션(duration) 및 인토네이션(intonation)을 포함할 수 있다. 음성 생성 장치(100)는 데이터베이스에서 추출한 두 특징과 유사한 복수의 후보 음성 유닛을 선택할 수 있다. 음성 생성 장치(100)는 입력 텍스트에 대응하는 음성 단위 별로 복수의 후보 음성 유닛을 선택할 수 있다. 이후에, 복수의 후보 음성 유닛은 비터비 검색의 대상이 된다.

[0046] 후보 음성 유닛이 미리 설정된 조건을 만족하지 않는 경우, 음성 생성 장치(100)는 새로운 음성 유닛을 생성할 수 있다. 음성 생성 장치(100)는 각 음성 단위에 대해 선택된 후보 음성 유닛의 코스트를 계산할 수 있다. 계산된 코스트가 임계값보다 큰 경우에, 음성 생성 장치(100)는 해당 음성 단위에 대응하는 새로운 음성 유닛을 생성할 수 있다. 예를 들어, "ㅏ"에 대응하여 선택된 음성 유닛의 코스트가 임계값보다 크다고 판단될 경우, 음성 생성 장치(100)는 "ㅏ"에 대응하는 새로운 음성 유닛을 생성할 수 있다.

[0047] 이처럼, 음성 생성 장치(100)는 각각의 음성 단위에 대응하여 코스트 조건을 만족하는 복수의 후보 음성 유닛을 선택하거나 생성할 수 있다. 최종적으로, 음성 생성 장치(100)는 모든 음성 단위에 대한 복수의 후보 음성 유닛에 대해 비터비 검색을 수행하여 최적의 음성 시퀀스를 획득할 수 있다. 음성 생성 장치(100)는 음성 시퀀스를 자연스럽게 결합하여 "안녕하세요"라는 출력 음성(103)을 스피커를 통하여 출력할 수 있다.

[0049] 도 2는 일 실시예에 따른 음성 생성 방법의 동작을 도시한 순서도이다.

[0050] 일 실시예에 따르면, 단계(201)에서, 음성 생성 장치(100)는 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득한다.

[0051] 일 실시예에 따르면, 단계(203)에서, 음성 생성 장치(100)는 링귀스틱 특징 및 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정한다. 데이터베이스에는 미리 다양한 링귀스틱 특징 및 프로소디 특징에 대응하는 다수의 음성 유닛(voice unit)이 저장될 수 있다. 음성 생성 장치(100)는 음성 유닛을 선택할 때 각 음성 유닛의 코스트(cost)를 계산하고 비터비(Viterbi) 검색을 수행할 수 있다.

[0052] 일 실시예에 따르면, 단계(205)에서, 음성 생성 장치(100)는 링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성한다. 음성 생성 장치(100)는 가장 작은 코스트를 가지는 음성 유닛으로부터 스타일 특징(style feature)과 컨텐츠 특징(content featrure)을 추출하고 이들을 합성하여 음성 유닛을 새로 생성할 수 있다. 다른 예로, 음성 생성 장치(100)는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 이용하여 음성 유닛을 생성할 수도 있다.

[0053] 일 실시예에 따르면, 단계(207)에서, 음성 생성 장치(100)는 제2 후보 음성 유닛을 비터비 검색을 통해 결정된

음성 시퀀스와 연결하여 출력 음성을 출력한다. 음성 생성 장치(100)는 음성 시퀀스를 자연스럽게 결합하여 최종 출력 음성을 스피커 등을 통하여 출력할 수 있다.

[0054] 일 실시예에 따르면, 음성 생성 장치(100)는 제1 후보 음성 시퀀스를 구성하는 후보 음성 유닛의 코스트가 임계값보다 클 경우, 음성 유닛 생성기에 의해 복수의 후보 음성 유닛을 생성하여 메모리에 저장할 수 있다. 예를 들어, 음성 생성 장치(100)는 학습 과정에서 접하지 못한(unseen) 다양한 포네틱 트랜스크립션(phonetic transcription)을 이용하여 새로운 음성 유닛을 생성하고, 이를 데이터베이스에 저장할 수 있다. 이 경우, 추론(inference) 과정에서 새로운 음성 유닛을 생성할 확률이 줄어들기 때문에, 음성 생성 장치(100)는 더욱 빨리 출력 음성을 출력할 수 있다.

[0055] 다른 실시예에 따르면, 음성 생성 장치(100)는 입력 텍스트에 대응하는 각각의 음성 단위 마다 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다. 예를 들어, 음성 생성 장치(100)는 각각의 음성 단위에 대해 매번 새로운 음성 유닛을 생성하여 복수의 후보 음성 유닛 그룹에 포함시키고, 비터비 검색을 통해 최적의 음성 시퀀스를 찾을 수 있다.

[0056] 다른 실시예에 따르면, 음성 생성 장치(100)는 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다. 예를 들어, 음성 생성 장치(100)는 데이터베이스에서 검색된 후보 음성 유닛의 코스트가 미리 설정된 조건을 만족하지 못하는 경우에만 새로운 음성 유닛을 생성할 수도 있다.

[0058] 도 3은 일 실시예에 따른 음성 생성 방법의 동작을 도시한 흐름도이다.

[0059] 단계(301)에서, 음성 생성 장치(100)는 입력 텍스트를 입력 받을 수 있다. 입력 텍스트는 복수의 음성 단위로 구성될 수 있다. 입력 텍스트를 음성으로 변환한 포네틱 트랜스크립션의 분절 단위는 복수일 수 있다.

[0060] 단계(310)에서, 음성 생성 장치(100)는 입력 텍스트의 특징을 추출할 수 있다. 단계(311)에서, 음성 생성 장치(100)는 입력 텍스트로부터 링귀스틱 특징을 추출할 수 있다. 단계(313)에서, 음성 생성 장치(100)는 링귀스틱 특징을 기초로 프로소디 특징을 추출하거나 생성할 수 있다.

[0061] 단계(320)에서, 음성 생성 장치(100)는 후보 음성 유닛을 선택할 수 있다. 음성 생성 장치(100)는 다양한 포네틱 트랜스크립션에 대응하는 음성 유닛을 저장한 데이터베이스로부터 추출된 특징에 적합한 후보 음성 유닛을 선택할 수 있다.

[0062] 단계(330)에서, 음성 생성 장치(100)는 선택된 후보 음성 유닛을 기초로 비터비 검색을 수행하고, 수행 결과를 평가할 수 있다. 단계(331)에서, 음성 생성 장치(100)는 선택된 후보 음성 유닛의 코스트를 계산할 수 있다. 단계(331)에서, 음성 생성 장치(100)는 각각의 음성 단위에 대한 복수의 후보 음성 유닛에 대해 비터비 검색을 수행하여 최적의 음성 시퀀스를 결정할 수 있다. 예를 들어, 음성 생성 장치(100)는 코스트 계산을 통해 최적의 음성 시퀀스를 결정할 수 있다. 음성 생성 장치(100)는 음성 단위 각각의 음성 유닛의 코스트와 각 음성 유닛 간의 코스트를 종합하여 전체 코스트를 계산할 수 있다. 음성 생성 장치(100)는 전체 코스트가 가장 작은 조합을 음성 시퀀스로 결정할 수 있다.

[0063] 음성 시퀀스의 코스트가 미리 설정된 조건을 만족한 경우, 단계(340)에서, 음성 생성 장치(100)는 음성 시퀀스의 각 음성 유닛을 결합하여 자연스러운 출력 음성을 출력할 수 있다. 반면에, 음성 시퀀스의 코스트가 미리 설정된 조건을 만족하지 못한 경우, 단계(350)에서, 음성 생성 장치(100)는 음성 유닛 생성기를 통해 새로운 음성 유닛을 생성할 수 있다. 음성 생성 장치(100)는 새로운 음성 유닛을 포함한 새로운 후보군을 대상으로 비터비 검색을 수행하여 코스트가 더욱 낮은 음성 시퀀스를 결정할 수 있다.

[0065] 도 4는 일 실시예에 따른 음성 생성 방법의 일례를 도시한 흐름도이다.

[0066] 음성 생성 장치(100)는 링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성한다. 음성 시퀀스의 코스트가 미리 설정된 조건을 만족하지 못한 경우, 단계(350)에서, 음성 생성 장치(100)는 음성 유닛 생성기를 통해 새로운 음성 유닛을 생성할 수 있다. 음성 생성 장치(100)는 새로운 음성 유닛을 포함한 새로운 후보군을 대상으로 비터비 검색을 수행하여 코스트가 더욱 낮은 음성 시퀀스를 결정할 수 있다.

[0067] 단계(453)에서, 음성 생성 장치(100)는 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제3 후보 음성 유닛으로부터 스타일(style) 특징을 추출할 수 있다. 단계(455)에서, 음성 생성 장치(100)는 복수의 후보 음성 유닛 중에서 제2 후보 음성 유닛과 상이한 제4 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출할 수

있다. 단계(451)에서, 음성 생성 장치(100)는 링귀스틱 특징, 스타일 특징 및 컨텐츠 특징을 기초로 제2 후보 음성 유닛을 생성할 수 있다. 음성 생성 장치(100)는 새로운 음성 유닛을 포함한 새로운 후보군을 대상으로 비터비 검색을 수행하여 코스트가 더욱 낮은 음성 시퀀스를 결정할 수 있다.

[0069] 도 5는 다른 실시예에 따른 음성 생성 방법의 일례를 도시한 흐름도이다.

[0070] 음성 생성 장치(100)는 링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성한다. 음성 시퀀스의 코스트가 미리 설정된 조건을 만족하지 못한 경우, 단계(350)에서, 음성 생성 장치(100)는 음성 유닛 생성기를 통해 새로운 음성 유닛을 생성할 수 있다.

[0071] 음성 생성 장치(100)는 링귀스틱 특징 및 프로소디 특징을 기초로 제2 후보 음성 유닛을 생성할 수 있다. 여기서, 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 포함할 수 있다. 음성 유닛 생성기는 생성자와 감별자의 경쟁적인 학습 과정을 통해 학습된 뉴럴 네트워크를 포함할 수 있다.

[0073] 도 6은 일 실시예에 따른 음성 유닛 생성기의 학습 방법의 동작을 도시한 순서도이다.

[0074] 일 실시예에 따르면, 단계(601)에서, 음성 유닛 생성기 학습 장치는 프로세서에 의해, 학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득한다. 학습 텍스트는 다양한 종류의 문장과 그에 대응하는 정답 음성의 쌍으로 구성될 수 있다.

[0075] 일 실시예에 따르면, 단계(603)에서, 음성 유닛 생성기 학습 장치는 링귀스틱 특징 및 프로소디 특징을 기초로 코스트 함수를 이용한 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정한다.

[0076] 일 실시예에 따르면, 단계(605)에서, 음성 유닛 생성기 학습 장치는 링귀스틱 특징 또는 프로소디 특징 및 제1 후보 음성 유닛의 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성한다.

[0077] 일 실시예에 따르면, 단계(607)에서, 음성 유닛 생성기 학습 장치는 제1 후보 음성 유닛 및 제2 후보 음성 유닛을 기초로 제2 후보 음성 유닛에 대응하는 손실값을 계산한다.

[0078] 일 실시예에 따르면, 단계(609)에서, 음성 유닛 생성기 학습 장치는 손실값을 기초로 음성 유닛 생성기의 파라미터를 갱신한다.

[0079] 코스트 함수가 미분 가능한 경우, 음성 유닛 생성기 학습 장치는 코스트 함수를 이용하여 손실값을 계산할 수 있다. 예를 들어, 음성 유닛 생성기 학습 장치는 코스트 함수가 미분 가능할 경우에 코스트 함수의 그레디언트(gradient)를 계산하고, 계산 결과를 기초로 역전파 기법을 이용하여 음성 유닛 생성기의 파라미터를 조정할 수 있다.

[0080] 예를 들어, 음성 유닛 생성기는 가장 작은 코스트를 가지는 음성 유닛으로부터 스타일 특징(style feature)을 추출하는 스타일 추출기와 컨텐츠 특징(content featrure)을 추출하는 컨텐츠 추출기를 포함할 수 있다. 이 경우, 음성 유닛 생성기 학습 장치는 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제2 후보 음성 유닛으로부터 스타일(style) 특징을 추출하는 스타일 추출기의 파라미터 및 복수의 후보 음성 유닛 중에서 제2 후보 음성 유닛과 상이한 제3 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출하는 특징 추출기를 포함하는 음성 유닛 생성기의 파라미터를 갱신할 수 있다.

[0081] 다른 예로, 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)를 포함할 수도 있다. 이 경우, 음성 유닛 생성기 학습 장치는 음성 유닛 생성기에 포함된 생성자(Generator) 및 감별자(Discriminator)의 파라미터를 갱신할 수 있다.

[0082] 생성자와 감별자는 링귀스틱 특징 및 프로소디 특징을 입력 받을 수 있다. 이러한 특징들을 통해 음성 유닛 생성기는 생성하고 감별해야 할 음성 유닛을 특정할 수 있다. 음성 유닛 생성기는 입력 받은 특징을 기초로 기존 음성 유닛으로부터 배운 지식을 이용하여 새로운 음성 유닛을 생성할 수 있다. 생성자는 입력 받은 특징과 함께 노이즈 데이터를 입력 받아 새로운 음성 유닛을 생성할 수 있다. 감별자는 코스트 계산과 비터비 검색을 통해 선택된 음성 유닛과 생성자가 생성한 음성 유닛 중에서 어느 것이 생성된 것인지를 판단할 수 있다. 여기서, 코스트 계산과 비터비 검색을 통해 선택된 최소 코스트를 가지는 음성 유닛은 진짜(Real)인 것으로 설정되고, 생성된 음성 유닛은 가짜(fake)인 것으로 설정된다. 매 단계마다 손실값이 계산되고 생성자와 감별자의 파라미터가 갱신될 수 있다. 생성자는 감별자가 가짜를 선택하도록 진짜 같은 음성 유닛을 생성하도록 학습되고, 감별자는 진짜인 음성 유닛을 선택하도록 학습될 수 있다. 이처럼, 경쟁적인 학습 과정을 통해 생성자는 보다 진짜 같은 음성 유닛을 생성할 수 있다. 추론 과정에서는 학습된 생성자만 사용될 수 있다.

[0084] 도 7은 일 실시예에 따른 음성 생성 장치의 구성을 도시한 도면이다.

[0085] 일 실시예에 따르면, 음성 생성 장치(700)는 적어도 하나의 프로세서(701) 및 음성 유닛 생성기를 저장하는 메모리(703)를 포함한다. 프로세서(701)는 입력 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득할 수 있다. 프로세서(701)는 링귀스틱 특징 및 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정할 수 있다. 프로세서(701)는 링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다. 프로세서(701)는 제2 후보 음성 유닛을 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력할 수 있다.

[0086] 일 실시예에 따르면, 프로세서(701)는 음성 유닛 생성기에 의해, 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제3 후보 음성 유닛으로부터 스타일(style) 특징을 추출할 수 있다. 프로세서(701)는 음성 유닛 생성기를 통해 복수의 후보 음성 유닛 중에서 제2 후보 음성 유닛과 상이한 제4 후보 음성 유닛으로부터 컨텐츠(content) 특징을 추출할 수 있다. 프로세서(701)는 음성 유닛 생성기에 의해, 링귀스틱 특징, 스타일 특징 및 컨텐츠 특징을 기초로 제2 후보 음성 유닛을 생성할 수 있다.

[0087] 다른 실시예에 따르면, 프로세서(701)는 음성 유닛 생성기에 의해, 링귀스틱 특징 및 프로소디 특징을 기초로 제2 후보 음성 유닛을 생성할 수 있다. 여기서, 음성 유닛 생성기는 생성적 적대 신경망(Generative Adversarial Network, GAN)을 포함할 수 있다.

[0088] 일 실시예에 따르면, 프로세서(701)는 제1 후보 음성 시퀀스를 구성하는 후보 음성 유닛의 코스트가 임계값보다 클 경우, 음성 유닛 생성기에 의해 복수의 후보 음성 유닛을 생성하여 메모리에 저장할 수 있다. 다른 실시예에 따르면, 프로세서(701)는 입력 텍스트에 대응하는 각각의 음성 단위 마다 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다. 다른 실시예에 따르면, 프로세서(701)는 제1 후보 음성 유닛의 코스트가 임계값보다 클 경우, 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다.

[0090] 도 8은 일 실시예에 따른 음성 유닛 생성기의 학습 장치의 구성을 도시한 도면이다.

[0091] 일 실시예에 따르면, 음성 유닛 생성기의 학습 장치(800)는 적어도 하나의 프로세서(801) 및 음성 유닛 생성기를 저장하는 메모리(803)를 포함한다. 프로세서(801)는 학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득할 수 있다. 프로세서(801)는 링귀스틱 특징 및 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정할 수 있다. 프로세서(801)는 링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성할 수 있다. 프로세서(801)는 제1 후보 음성 유닛 및 제2 후보 음성 유닛을 기초로 제2 후보 음성 유닛에 대응하는 손실값을 계산할 수 있다. 프로세서(801)는 손실값을 기초로 음성 유닛 생성기의 파라미터를 갱신할 수 있다.

[0092] 코스트 함수가 미분 가능한 경우, 음성 유닛 생성기 학습 장치(800)는 코스트 함수를 이용하여 손실값을 계산할 수 있다. 예를 들어, 음성 유닛 생성기 학습 장치(800)는 코스트 함수가 미분 가능할 경우에 코스트 함수의 그레디언트(gradient)를 계산하고, 계산 결과를 기초로 역전파 기법을 이용하여 음성 유닛 생성기의 파라미터를 조정할 수 있다.

[0093] 일 실시예에 따르면, 프로세서(801)는 복수의 후보 음성 유닛 중에서 가장 작은 코스트를 가지는 제2 후보 음성 유닛으로부터 스타일(style) 특징을 스타일 추출기의 파라미터 및 복수의 후보 음성 유닛 중에서 제2 후보 음성 유닛과 상이한 제3 후보 음성 유닛으로부터 컨텐츠(content) 특징을 특징 추출기를 포함하는 음성 유닛 생성기의 파라미터를 갱신할 수 있다.

[0094] 다른 실시예에 따르면, 프로세서(801)는 음성 유닛 생성기에 포함된 생성자 (Generator) 및 감별자(Discriminator)의 파라미터를 갱신할 수 있다.

[0096] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설

명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

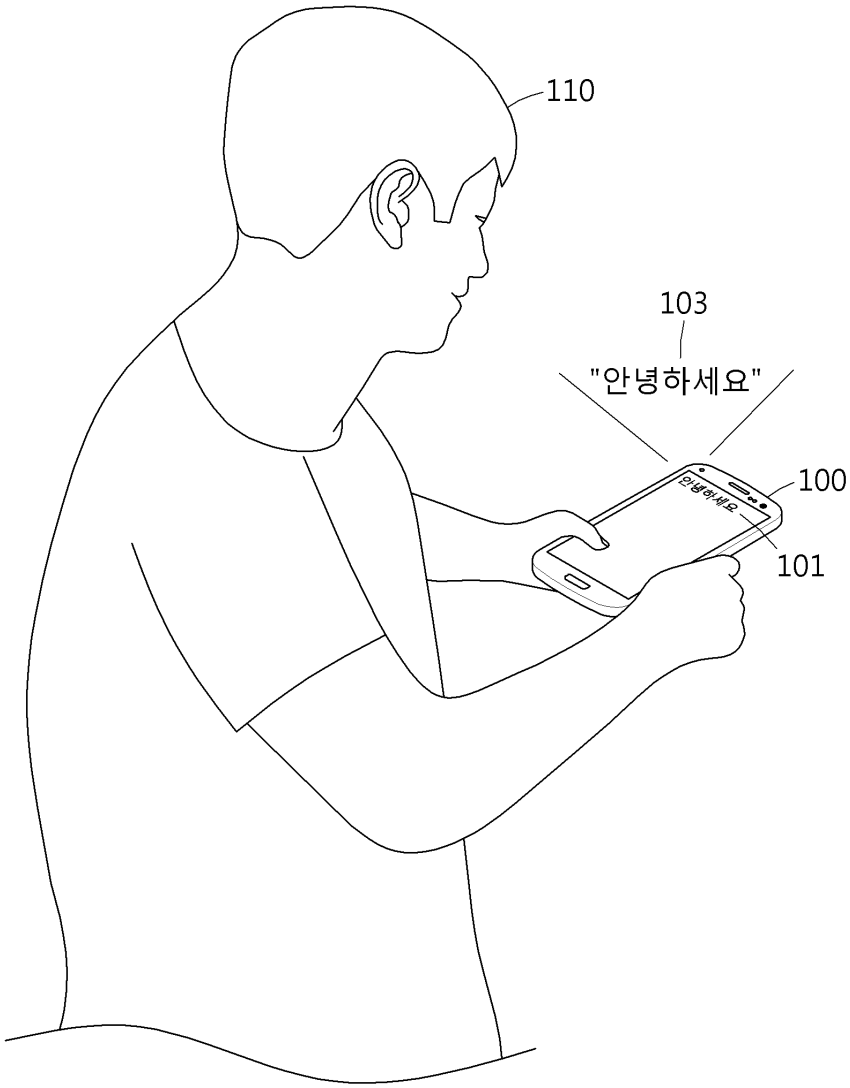
[0097] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0098] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

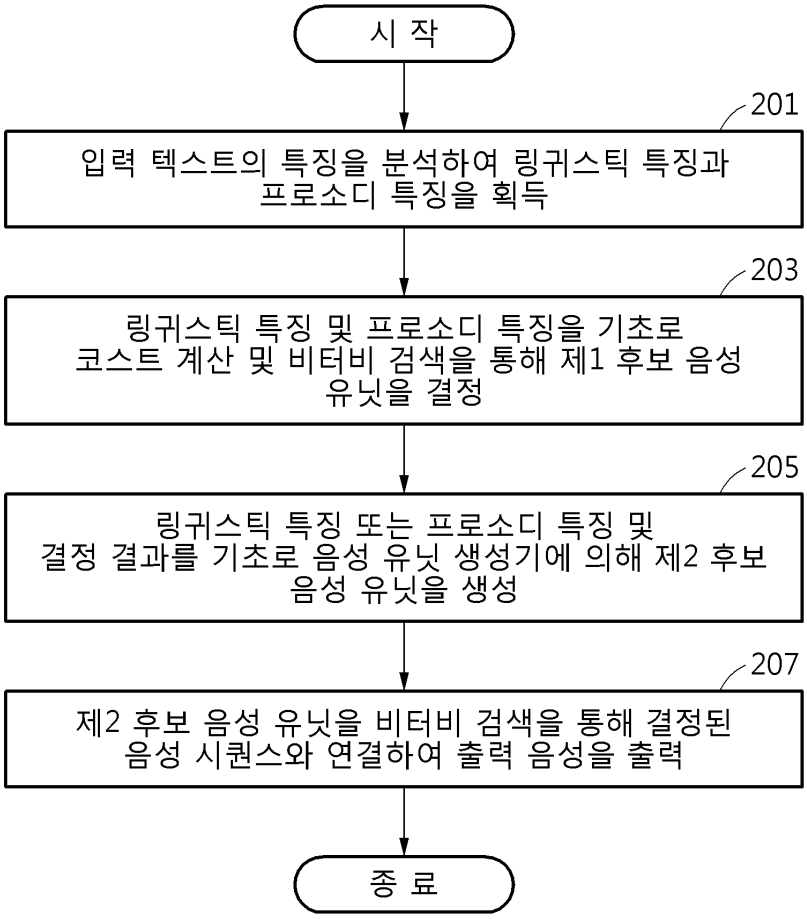
[0099] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

도면

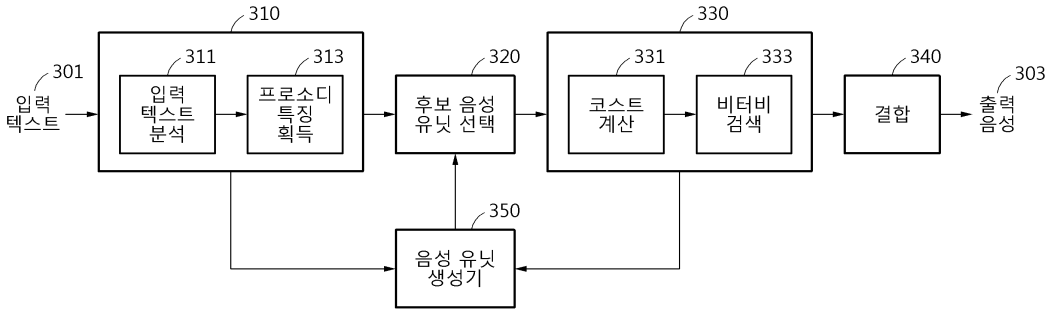
도면1



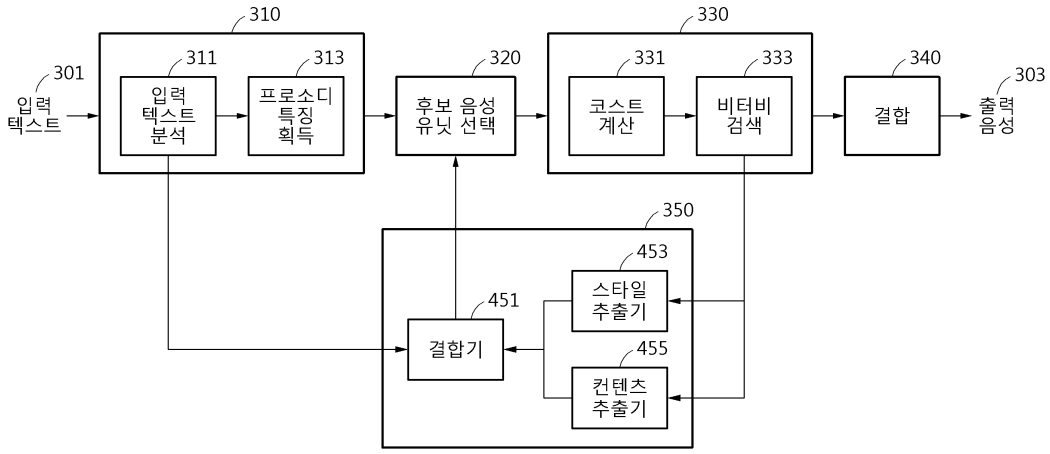
도면2


시 작
201
입력 텍스트의 특징을 분석하여 링귀스틱 특징과 프로소디 특징을 획득
203
링귀스틱 특징 및 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정
205
링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성
207
제2 후보 음성 유닛을 비터비 검색을 통해 결정된 음성 시퀀스와 연결하여 출력 음성을 출력
종 료

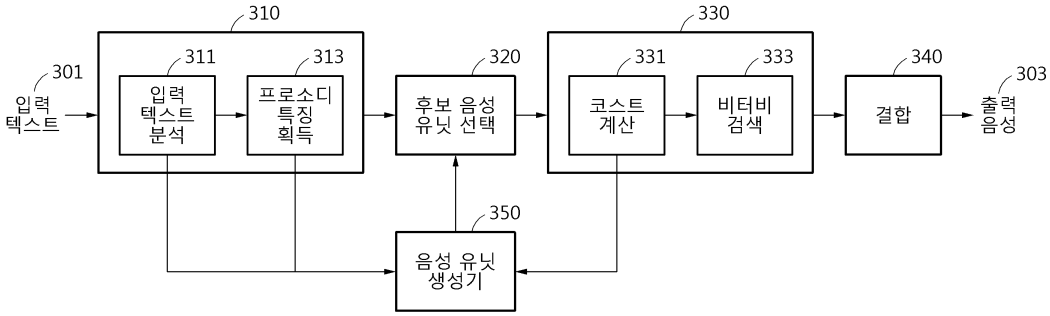
도면3


310
301
입력 텍스트
311
입력 텍스트 분석
313
프로소디 특징 획득
320
후보 음성 유닛 선택
330
331
코스트 계산
333
비터비 검색
340
결합
303
출력 음성
350
음성 유닛 생성기

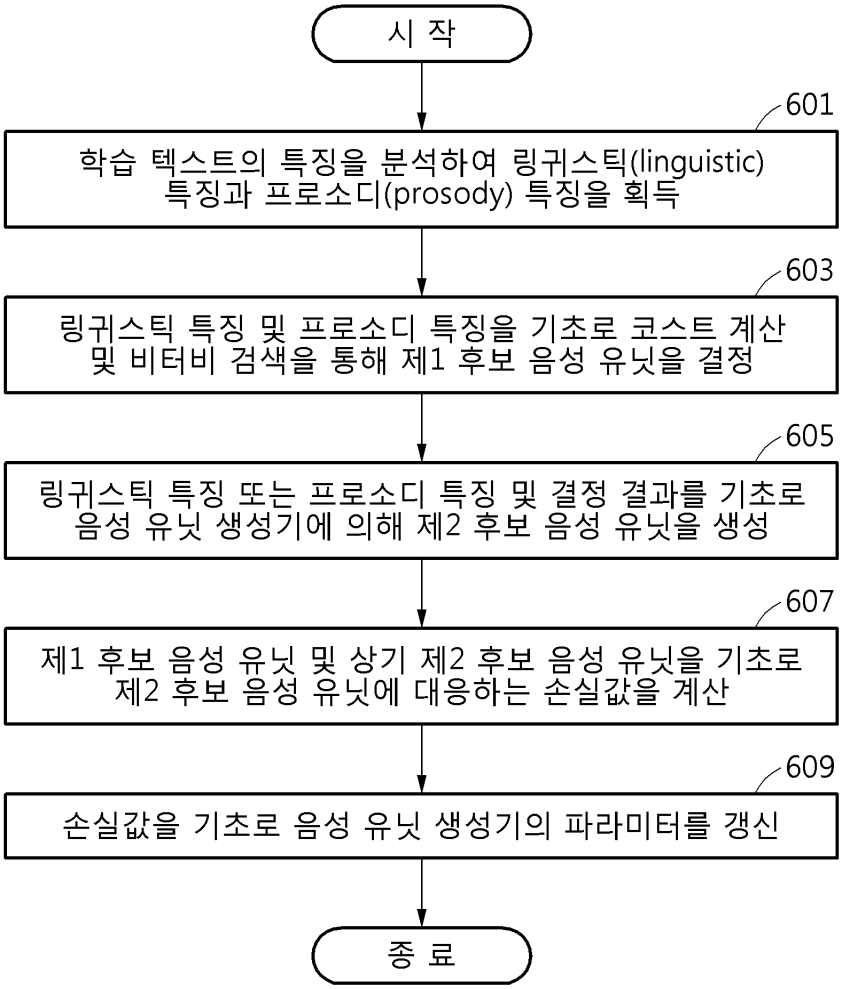
도면4


310
301
입력
텍스트
311
입력
텍스트
분석
313
프로소디
특징
획득
320
후보 음성
유닛 선택
330
코스트
계산
331
333
비터비
검색
340
결합
303
출력
음성
350
453
스타일
추출기
451
결합기
455
컨텐츠
추출기

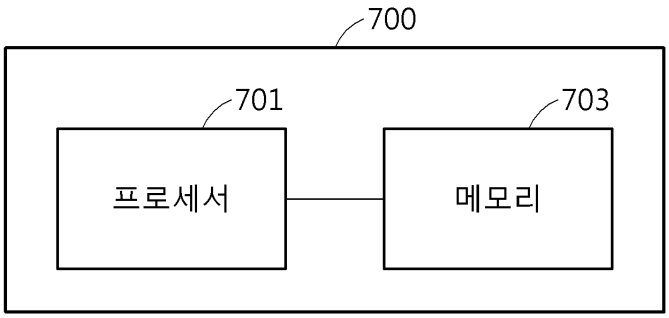
도면5


310
301
입력
텍스트
311
입력
텍스트
분석
313
프로소디
특징
획득
320
후보 음성
유닛 선택
330
코스트
계산
331
333
비터비
검색
340
결합
303
출력
음성
350
음성 유닛
생성기

도면6


시 작
601
학습 텍스트의 특징을 분석하여 링귀스틱(linguistic) 특징과 프로소디(prosody) 특징을 획득
603
링귀스틱 특징 및 프로소디 특징을 기초로 코스트 계산 및 비터비 검색을 통해 제1 후보 음성 유닛을 결정
605
링귀스틱 특징 또는 프로소디 특징 및 결정 결과를 기초로 음성 유닛 생성기에 의해 제2 후보 음성 유닛을 생성
607
제1 후보 음성 유닛 및 상기 제2 후보 음성 유닛을 기초로 제2 후보 음성 유닛에 대응하는 손실값을 계산
609
손실값을 기초로 음성 유닛 생성기의 파라미터를 갱신
종 료

도면7


700
701
프로세서
703
메모리

도면8