



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0020082
(43) 공개일자 2024년02월14일

- | | |
|---|---|
| <p>(51) 국제특허분류(Int. Cl.)
 $G10L\ 13/033$ (2013.01) $G06N\ 20/00$ (2019.01)
 $G10L\ 13/047$ (2013.01) $G10L\ 21/0272$ (2013.01)
 $G10L\ 25/69$ (2013.01)</p> <p>(52) CPC특허분류
 $G10L\ 13/033$ (2013.01)
 $G06N\ 20/00$ (2021.08)</p> <p>(21) 출원번호 10-2022-0098093
 (22) 출원일자 2022년08월05일
 심사청구일자 2022년08월05일</p> | <p>(71) 출원인
 이화여자대학교 산학협력단
 서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)</p> <p>(72) 발명자
 도인실
 인천광역시 연수구 센트럴로 194, 201동 1703호(송도동, 더샵 센트럴파크2)</p> <p>최윤정
 서울특별시 노원구 중계로 233, 108동 704호(중계동, 청구3차아파트)
 (뒷면에 계속)</p> <p>(74) 대리인
 김연권</p> |
|---|---|

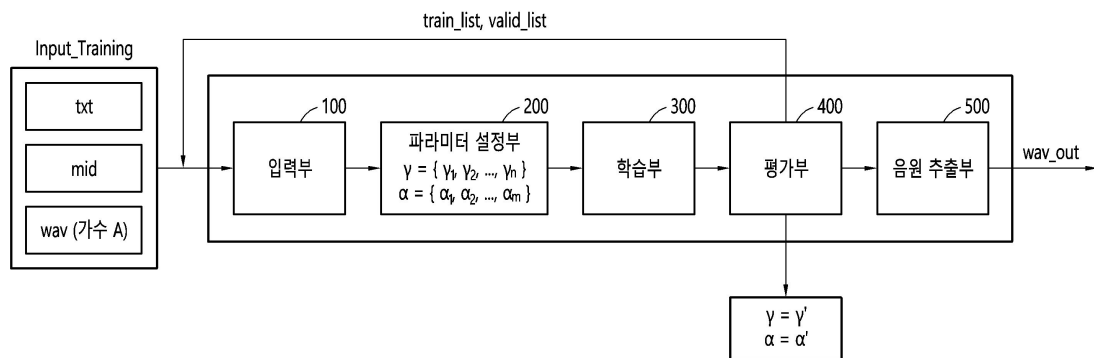
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 딥러닝 기반 가창 음성 합성 장치 및 이를 이용한 가창 음성 합성 방법

(57) 요약

가창 음성 합성 장치는 가사 데이터, 미디 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신하는 입력부, 상기 입력 데이터에 기초하여 감마값 및 알파값을 설정하는 파라미터 설정부, 설정된 상기 감마값 및 상기 알파값을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 학습부, 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 평가부, 및 상기 가창 음성 데이터에 기초하여 최종 음원을 추출하는 음원 추출부를 포함할 수 있다. 상기 감마값은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터이고, 상기 알파값은 생성자 손실 함수에 포함된 L1 손실의 가중치일 수 있다. 상기 학습부는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 가창 음성 데이터를 합성할 수 있다.

대표도



(52) CPC특허분류

G10L 13/047 (2013.01)

G10L 21/0272 (2013.01)

G10L 25/69 (2013.01)

(72) 발명자

박지현

경기도 수원시 장안구 이목로 24, 114동 2304호(정자동, 수원 SK SKY VIEW)

김민애

경기도 평택시 안중읍 안현로서9길 164-20, 102동 601호(화현마을우림필유)

허가빈

경상남도 창원시 의창구 하남천서길21번길 8-14, 1107호(도계동, 강남훼미리아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711154280

과제번호 2020R1A2C1006497

부처명 과학기술정보통신부

과제관리(전문)기관명 한국연구재단

연구사업명 이공분야기초연구사업(통합)-중견연구자지원사업-(유형1-1)중견연구

연구과제명 UAV를 활용한 모바일 IoT 환경에서 효율적이고 안전한 노드 및 데이터 프라이버시

보장 기법

기 여 율 1/1

과제수행기관명 이화여자대학교

연구기간 2020.03.01 ~ 2023.02.28

명세서

청구범위

청구항 1

가사 데이터, 미디 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신하는 입력부;

상기 입력 데이터에 기초하여 감마값 및 알파값을 설정하는 파라미터 설정부;

설정된 상기 감마값 및 상기 알파값을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 학습부;

상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 평가부; 및

상기 가창 음성 데이터에 기초하여 최종 음원을 추출하는 음원 추출부를 포함하고,

상기 감마값은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터이고,

상기 알파값은 생성자 손실 함수에 포함된 L1 손실의 가중치이고,

상기 학습부는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 가창 음성 데이터를 합성하는 것을 특징으로 하는,

가창 음성 합성 장치.

청구항 2

제1항에 있어서,

상기 생성자 손실 함수는 아래 [수식1]을 이용하여 산출되는 것을 특징으로 하는,

가창 음성 합성 장치.

[수식1]

$$L_G = L(G(x)) + \alpha \cdot |y - G(x)|$$

(여기서, L_G 는 생성자 손실 함수이고, $G(x)$ 는 가짜 샘플이고, y 는 실제 샘플이고, α 는 알파값을 나타낸다.)

청구항 3

제1항에 있어서,

상기 평가부는,

상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 상기 감마값 및 상기 알파값을 최적화 파라미터로 저장하는 것을 특징으로 하는,

가창 음성 합성 장치.

청구항 4

제1항에 있어서,

상기 평가부는,

상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 작은 경우, 학습 리스트 데이터 및 평가 리스트 데이터를 상기 입력부의 입력으로 피드백하는 것을 특징으로 하는,

가창 음성 합성 장치.

청구항 5

제4항에 있어서,

상기 파라미터 설정부는,

상기 가창 음성 평가 점수가 상기 기준 평가 점수보다 작은 경우, 상기 학습 리스트 데이터 및 상기 평가 리스트 데이터에 기초하여 상기 감마값 및 상기 알파값을 재설정하는 것을 특징으로 하는,

가창 음성 합성 장치.

청구항 6

제1항에 있어서,

상기 음원 추출부는,

상기 평가부가 최적화 파라미터로 저장하는 상기 감마값 및 상기 알파값이 적용된 상기 가창 음성 데이터에 대하여 음원 튜닝을 수행함으로써 상기 최종 음원을 추출하는 것을 특징으로 하는,

가창 음성 합성 장치.

청구항 7

가사 데이터, 미디 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신하는 단계;

상기 입력 데이터에 기초하여 감마값 및 알파값을 설정하는 단계;

설정된 상기 감마값 및 상기 알파값을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 단계;

상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계; 및

상기 가창 음성 데이터에 기초하여 최종 음원을 추출하는 단계를 포함하고,

상기 감마값은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터이고,

상기 알파값은 생성자 손실 함수에 포함된 L1 손실의 가중치이고,

상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 단계는, BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 가창 음성 데이터를 합성하는 것을 특징으로 하는,

가창 음성 합성 방법.

청구항 8

제7항에 있어서,

상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계는,

상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 상기 감마값 및 상기 알파값을 최적화 파라미터로 저장하는 것을 특징으로 하는,

가창 음성 합성 방법.

청구항 9

제7항에 있어서,

상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계는,

상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 작은 경우, 학습 리스트 데이터 및 평가 리스트 데이터를 상기 입력부의 입력으로 피드백하는 것을 특징으로 하는,

가창 음성 합성 방법.

발명의 설명

기술 분야

[0001] 본 발명은 가창 음성 합성에 관한 것으로, 보다 상세하게는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 학습 모델을 이용한 딥러닝 기반 가창 음성 합성 장치 및 이를 이용한 가창 음성 합성 방법에 관한 것이다.

배경 기술

[0002] 오랜 기간동안 진행된 TTS(Text-To-Speech) 합성 모델을 바탕으로 고수준의 음성 합성 시스템들이 개발되고 있다. 더 나아가 음정과 감정을 포괄하여 가창 데이터를 합성해내는 SVS(Singing Voice Synthesis)로 발전하고 있으며 메타버스와 가상 엔터테이너에 대한 관심이 급증하면서 다양한 SVS모델에 대한 연구 또한 활발하다.

[0003] SVS란 가사(text)와 노래하는 음성(voice), 그리고 멜로디 정보가 담겨있는 미디(midi) 데이터를 입력으로 하여 자연스럽게 노래를 합성하는 시스템이다. 최근에는 AI기반 지능형 디바이스들의 응용시나리오 수준이 높아짐에 따라 SVS 모델에도 심층신경망이 적용되고 있다.

[0004] 예를 들어, 딥러닝이나 종단기술(end-to-end)을 활용하는 고성능 알고리즘 등 기존 이미지 합성에 주로 적용되었던 DNN, LSTM, GAN 등이 SVS 모델에 적용되고 있다.

[0005] 대표적으로, GAN은 주요한 SVS 생성 모델들 중 하나이다. GAN은 모델이 수렴에 도달하면 생성자는 실제 샘플과 거의 구분되지 않는 샘플을 생성해 낼 수 있다.

[0006] 그러나, 종래 기술의 GAN을 이용하는 경우, 어느 순간 생성자가 의미없이 계속 같은 모양의 생성물을 출력해내는 모드붕괴(mode collapse) 현상이 발생할 수 있다. 일례로 판별자를 속이는 데만 최적화된 생성자와 지역 최적화(local minima)에 빠진 판별자 하에서는 학습 불안정 및 학습 불균형에 따른 GAN의 학습 성능 저하가 발생할 수 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국공개특허 제 10-2021-0155520호 “복수의 가수의 가창음성 합성/변조 장치 및 방법”

발명의 내용

해결하려는 과제

[0008] 본 발명의 일 목적은 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 학습 모델을 적용하고, 생성자 손실의 과적합 문제를 최소화함으로써, 실제 사람이 부른 것 같은 자연스러운 음원을 만드는 가창 음성 합성 장치를 제공하는 것이다.

[0009] 본 발명의 다른 목적은 본 발명의 일 목적은 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 학습 모델을 적용하고, 생성자 손실의 과적합 문제를 최소화함으로써, 실제 사람이 부른 것 같은 자연스러운 음원을 만드는 가창 음성 합성 방법을 제공하는 것이다.

[0010] 다만, 본 발명이 해결하고자 하는 과제는 상기 언급된 과제에 한정되는 것이 아니며, 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위에서 다양하게 확장될 수 있을 것이다.

과제의 해결 수단

[0011] 본 발명의 일 목적을 달성하기 위하여, 본 발명의 실시예들에 따른 가창 음성 합성 장치는 가사 데이터, 미디 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신하는 입력부, 상기 입력 데이터에 기초하여 감마값 및 알파값을 설정하는 파라미터 설정부, 설정된 상기 감마값 및 상기 알파값을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 학습부, 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 평가부, 및 상기 가창 음성 데이터에 기초하여 최종 음원을 추출하는 음원 추출부를 포함

할 수 있다. 상기 감마값은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터이고, 상기 알파값은 생성자 손실 함수에 포함된 L1 손실의 가중치일 수 있다. 상기 학습부는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 가창 음성 데이터를 합성할 수 있다.

[0012] 일 실시예에서, 상기 생성자 손실 함수는 아래 [수식1]을 이용하여 산출될 수 있다.

[0013] [수식1]

$$L_G = L(G(x)) + \alpha \cdot |y - G(x)|$$

[0015] (여기서, L_G 는 생성자 손실 함수이고, $G(x)$ 는 가짜 샘플이고, y 는 실제 샘플이고, α 는 알파값을 나타낸다.)

[0016] 일 실시예에서, 상기 평가부는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 상기 감마값 및 상기 알파값을 최적화 파라미터로 저장할 수 있다.

[0017] 일 실시예에서, 상기 평가부는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 작은 경우, 학습 리스트 데이터 및 평가 리스트 데이터를 상기 입력부의 입력으로 피드백할 수 있다.

[0018] 일 실시예에서, 상기 파라미터 설정부는 상기 가창 음성 평가 점수가 상기 기준 평가 점수보다 작은 경우, 상기 학습 리스트 데이터 및 상기 평가 리스트 데이터에 기초하여 상기 감마값 및 상기 알파값을 재설정할 수 있다.

[0019] 일 실시예에서, 상기 음원 추출부는 상기 평가부가 최적화 파라미터로 저장하는 상기 감마값 및 상기 알파값이 적용된 상기 가창 음성 데이터에 대하여 음원 튜닝을 수행함으로써 상기 최종 음원을 추출할 수 있다.

[0020] 본 발명의 다른 목적을 달성하기 위하여, 본 발명의 실시예들에 따른 가창 음성 합성 방법은 가사 데이터, 미디어 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신하는 단계, 상기 입력 데이터에 기초하여 감마값 및 알파값을 설정하는 단계, 설정된 상기 감마값 및 상기 알파값을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 단계, 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계, 및 상기 가창 음성 데이터에 기초하여 최종 음원을 추출하는 단계를 포함할 수 있다. 상기 감마값은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터이고, 상기 알파값은 생성자 손실 함수에 포함된 L1 손실의 가중치일 수 있다. 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 단계는, BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 가창 음성 데이터를 합성할 수 있다.

[0021] 일 실시예에서, 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 상기 감마값 및 상기 알파값을 최적화 파라미터로 저장할 수 있다.

[0022] 일 실시예에서, 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 작은 경우, 학습 리스트 데이터 및 평가 리스트 데이터를 상기 입력부의 입력으로 피드백할 수 있다.

발명의 효과

[0023] 본 발명의 실시예들에 따른 가창 음성 합성 장치는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 가창 음성 데이터를 학습할 수 있다.

[0024] BEGAN 기반 학습 모델은 오토 인코더 아키텍처를 사용하여 손실 분포를 일치시킬 수 있다. 또한, BEGAN 기반 학습 모델은 판별자와 생성자 사이의 균형을 조정해주는 평형 개념(equilibrium concept)을 도입하여 생성 데이터의 다양성(diversity)과 품질(quality) 사이의 균형을 조정할 수 있다.

[0025] 따라서, 본 발명의 가창 음성 합성 장치는 BEGAN 기반 학습 모델을 이용하여 생성자 손실의 과적합 문제를 최소화함으로써, 실제 사람이 부른 것 같이 자연스러운 고품질 가창 음성 음원을 생성할 수 있다.

[0026] 다만, 본 발명의 효과는 상술한 효과에 한정되는 것이 아니며, 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위에서 다양하게 확장될 수 있을 것이다.

도면의 간단한 설명

- [0027] 도 1은 본 발명의 실시예들에 따른 가창 음성 합성 장치의 입력 및 출력을 나타내는 도면이다.
- 도 2는 학습 단계에서 도 1의 가창 음성 합성 장치의 구성 및 동작을 나타내는 블록도이다.
- 도 3은 파라미터 설정부가 감마값 및 알파값을 설정하는 일 예시를 나타내는 도면이다.
- 도 4는 학습부의 BEGAN 기반 학습 모델의 일 예시를 나타내는 도면이다.
- 도 5는 가창 음성 데이터 합성에 사용되는 손실 함수의 수식을 나타내는 도면이다.
- 도 6은 감마값 및 알파값의 임의의 조합을 나타내는 도표이다.
- 도 7은 도 6의 조합이 적용된 경우, 일부의 스펙트로그램 및 파형을 나타내는 도면이다.
- 도 8은 도 1의 가창 음성 합성 장치를 이용한 가창 음성 합성 방법을 나타내는 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0028] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.
- [0029] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0030] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만, 예를 들어 본 발명의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0031] 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. 구성요소들 간의 관계를 설명하는 표현들, 예를 들어 "~사이에"와 "바로~사이에" 또는 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.
- [0032] 본 명세서에서 사용한 용어는 단지 특정한 실시예들을 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 실시된 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.
- [0033] 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 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.
- [0034] 이하, 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.
- [0035] 도 1은 본 발명의 실시예들에 따른 가창 음성 합성 장치(10)의 입력 및 출력을 나타내는 도면이다.
- [0036] 도 1을 참조하면, 본 발명의 가창 음성 합성 장치(10)는 입력 데이터를 기초로 가창 음성을 합성하고, 가창 음성에 대한 음원을 출력할 수 있다.
- [0037] 예를 들어, 가창 음성 합성 장치(10)는 가사 데이터, 미디 데이터, 및 원곡 데이터(예컨대, 가수B)를 입력받고,

상기 원곡 데이터에 대한 타겟 가수(예컨대, 가수A)의 가창 음성 데이터를 합성할 수 있다.

- [0038] 가창 음성 합성 장치(10)는 상기 타겟 가수의 가창 음성 데이터에 기초하여 최종 음원을 추출할 수 있다.
- [0039] 구체적으로, 본 발명의 가창 음성 합성 장치(10)는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 가창 음성 데이터를 학습할 수 있다.
- [0040] 예를 들어, BEGAN 기반 학습 모델은 오토 인코더 아키텍처를 사용하여 손실 분포를 일치시킬 수 있다.
- [0041] 또한, BEGAN 기반 학습 모델은 판별자와 생성자 사이의 균형을 조정해주는 평형 개념(equilibrium concept)을 도입하여 생성 데이터의 다양성(diversity)과 품질(quality) 사이의 균형을 조정할 수 있다.
- [0042] 따라서, 본 발명의 가창 음성 합성 장치(10)는 BEGAN 기반 학습 모델을 이용함으로써, 종래기술의 가창 음성 합성 장치보다 상대적으로 높은 성능의 가창 음성 합성을 수행할 수 있다.
- [0043] 도 2는 학습 단계에서 도 1의 가창 음성 합성 장치(10)의 구성 및 동작을 나타내는 블록도이다.
- [0044] 도 2를 참조하면, 가창 음성 합성 장치(10)는 입력부(100), 파라미터 설정부(200), 학습부(300), 평가부(400), 및 음원 추출부(500)를 포함할 수 있다.
- [0045] 입력부(100)는 가사 데이터, 미디 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신할 수 있다.
- [0046] 예를 들어, 학습 단계에서, 입력부(100)는 학습 입력 데이터를 입력받을 수 있다.
- [0047] 학습 입력 데이터는 가사 데이터, 미디 데이터, 및 음원 데이터(예컨대, 가수A)를 포함할 수 있다.
- [0048] 가사 데이터는 노래의 가사에 대한 텍스트를 포함할 수 있다.
- [0049] 미디 데이터는 음고, 음의 길이 및 음의 세기에 대한 정보를 포함할 수 있다.
- [0050] 음원 데이터는 타겟 가수의 복수의 음원을 포함할 수 있다.
- [0051] 파라미터 설정부(200)는 상기 입력 데이터에 기초하여 감마값(γ) 및 알파값(α)을 설정할 수 있다.
- [0052] 예를 들어, 파라미터 설정부(200)는 입력부(100)로부터 상기 학습 입력 데이터를 수신하고, 상기 학습 입력 데이터에 상응하는 감마값(γ) 및 알파값(α)을 설정할 수 있다.
- [0053] 감마값(γ)은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터일 수 있다.
- [0054] 알파값(α)은 생성자 손실 함수에 포함된 L1 손실의 가중치일 수 있다.
- [0055] 한편, 감마값(γ) 및 알파값(α)의 설정 방법 및 최적화 방법에 대하여는 도 3 내지 5를 참조하여 상세히 설명한다.
- [0056] 학습부(300)는 설정된 상기 감마값(γ) 및 상기 알파값(α)을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성할 수 있다.
- [0057] 예를 들어, 학습부(300)는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 감마값(γ) 및 상기 알파값(α)이 적용된 상기 가창 음성 데이터를 합성할 수 있다.
- [0058] 평가부(400)는 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출할 수 있다.
- [0059] 평가부(400)는 상기 가창 음성 평가 점수를 기저장된 기준 평가 점수와 비교할 수 있다.
- [0060] 예를 들어, 상기 기준 평가 점수는 최종 음원을 기준으로 청취자가 음원에 대한 이질감을 느끼지 않는 최소 점수를 의미하며, 사용자가 설정 가능한 값일 수 있다.
- [0061] 상기 평가부(400)는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 작은 경우, 학습 리스트 데이터 및 평가 리스트 데이터를 상기 입력부(100)의 입력으로 피드백할 수 있다.
- [0062] 반대로, 상기 평가부(400)는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 상기 감마값(γ) 및 상기 알파값(α)을 최적화 파라미터로 저장할 수 있다.
- [0063] 가창 음성 평가 점수가 상기 기준 평가 점수보다 높은 경우, 평가부(400)는 가창 음성 데이터를 음원 추출부(500)에 출력할 수 있다.

- [0064] 음원 추출부(500)는 상기 가창 음성 데이터에 기초하여 최종 음원을 추출할 수 있다.
- [0065] 도 3은 파라미터 설정부(200)가 감마값(γ) 및 알파값(α)을 설정하는 일 예시를 나타내는 도면이다.
- [0066] 파라미터 설정부(200)는 입력부(100)로부터 상기 학습 입력 데이터를 수신하고, 가창 음성 합성에 적용되는 파라미터를 설정할 수 있다.
- [0067] 예를 들어, 감마값(γ)은 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 0 내지 1 사이의 값을 가지는 하이퍼 파라미터일 수 있다.
- [0068] 예를 들어, 알파값(α)은 생성자 손실 함수에 포함된 L1 손실의 가중치로서, 0 내지 1 사이의 값을 가질 수 있다.
- [0069] 파라미터 설정부(200)는 n개의 감마값($\gamma_1, \gamma_2, \dots, \gamma_n$) 및 m개의 알파값($\alpha_1, \alpha_2, \dots, \alpha_m$) 중 가창 음성 합성에 적용되는 감마값(γ) 및 알파값(α)을 설정할 수 있다.
- [0070] 예를 들어, 도 3에서 보듯이, 파라미터 설정부(200)는 $n * m$ 의 감마-알파 그리드를 만들고, 감마값(γ) 및 알파값(α)을 적절한 간격으로 배열하여 감마값(γ) 및 알파값(α)의 파라미터 조합을 조정할 수 있다.
- [0071] 파라미터 설정부(200)는 감마값(γ) 및 알파값(α)의 조합을 조정하고, 감마값(γ) 및 알파값(α)을 포함하는 파라미터를 학습부(300)에 출력할 수 있다.
- [0072] 도 4는 학습부(300)의 BEGAN 기반 학습 모델의 일 예시를 나타내는 도면이다.
- [0073] 학습부(300)는 BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 감마값(γ) 및 상기 알파값(α)이 적용된 상기 가창 음성 데이터를 합성할 수 있다.
- [0074] 도 4를 참조하면, BEGAN 기반 학습 모델은 도메인 인코더(Domain Encoder), 도메인 디코더(Domain Decoder), U-NET, 및 판별자(Discriminator)를 포함할 수 있다.
- [0075] 일 실시예에서, BEGAN 기반 학습 모델은 판별자로 오토 인코더(auto-encoder)를 사용할 수 있다.
- [0076] 예를 들어, 판별자는 오토 인코더(auto-encoder)를 사용하므로, 이미지를 복원하는 역할, 및 진짜 이미지와 가짜 이미지를 구별하는 역할을 각각 수행할 수 있다.
- [0077] 판별자는 안정적인 학습을 위하여 상기 두 가지 역할 간의 균형을 맞추기 위한 하이퍼 파라미터인 감마값(γ)이 적용될 수 있다.
- [0078] 예를 들어, 감마값(γ)은 아래 [수식]과 같이 나타낼 수 있다.
- [0079] [수식]
- [0080]
$$\gamma = E[L(G(x))]/E[L(y)]$$
- [0081] 여기서, γ 는 감마값(γ)이고, L은 오토 인코더의 재구성 에러이고, G(x)는 가짜 샘플이고, y는 실제 샘플일 수 있다.
- [0082] 예를 들어, 감마값(γ)이 작을수록(예컨대, 감마값(γ)이 0에 가까워질수록) 판별자가 실제 샘플 y를 오토 인코딩하는데 더 초점을 두기 때문에 생성된 샘플의 품질(quality)은 높아지지만 다양성(diversity)이 낮아질 수 있다.
- [0083] 반대로, 감마값(γ)이 높을수록(예컨대, 감마값(γ)이 1에 가까워질수록) 샘플의 다양성(diversity)이 높아지지만 샘플의 품질(quality)은 낮아질 수 있다.
- [0084] 도 5는 가창 음성 데이터 합성에 사용되는 손실 함수의 수식을 나타내는 도면이다.
- [0085] 도 5를 참조하면, 가창 음성 데이터 합성에 사용되는 손실 함수는 판별자 손실 함수 및 생성자 손실 함수를 포함할 수 있다. 또한, 손실 함수는 손실 조절 인자를 더 포함할 수 있다.
- [0086] 예를 들어, 판별자 손실 함수는 아래 [수식]과 같이 나타낼 수 있다.

- [0087] [수식]
- [0088]
$$L_D = L(y) - k_t \cdot L(G(x))$$
- [0089] 여기서, L_0 는 판별자 손실 함수이고, y 는 실제 샘플이고, $G(x)$ 는 가짜 샘플이고, k 는 손실 조절 인자를 나타낸다.
- [0090] 예를 들어, 생성자 손실 함수는 아래 [수식]과 같이 나타낼 수 있다.
- [0091] [수식]
- [0092]
$$L_G = L(G(x)) + \alpha \cdot |y - G(x)|$$
- [0093] 여기서, L_G 는 생성자 손실 함수이고, $G(x)$ 는 가짜 샘플이고, y 는 실제 샘플이고, α 는 알파값을 나타낸다.
- [0094] 예를 들어, 손실 조절 인자는 아래 [수식]과 같이 나타낼 수 있다.
- [0095] [수식]
- [0096]
$$k_{t+1} = k_t + \lambda_k(\gamma L(y) - L(G(x)))$$
- [0097] 여기서, k 는 손실 조절 인자이고, y 는 실제 샘플이고, $G(x)$ 는 가짜 샘플이고, γ 는 감마값이고, λ 는 학습률을 나타낼 수 있다.
- [0098] 예를 들어, 손실 조절 인자(k)는 감마값의 크기에 따라 $L(G(x))$ 를 얼마나 강조할 것인지 여부가 조절될 수 있다.
- [0099] 생성자 손실 함수(L_G)는 실제값(ground truth)와 생성물의 스펙트로그램(generated spectrogram)간의 픽셀 분포를 반영하기 위해 L1 손실을 포함할 수 있다.
- [0100] 다만, 생성자 손실 함수(L_G)에 L1 손실이 포함되는 경우, 감마값(γ)의 생성된 데이터의 다양성과 품질 사이의 균형을 조절하는 역할이 제한되는 문제가 발생할 수 있다.
- [0101] 예를 들어, 생성자 손실 함수(L_G)에 L1 손실이 포함되는 경우, 과적합(Over-fitting)에 의해 감마값(γ)이 낮아져도 특정 지점에서는 생성물의 품질이 높아지지 않는 현상이 발생할 수 있다.
- [0102] 이러한 문제를 해결하기 위하여, 본 발명의 생성자 손실 함수(L_G)는 L1 손실의 가중치 역할을 수행하는 알파값(α)을 포함할 수 있다.
- [0103] 예를 들어, 학습부(300)는 BEGAN 기반 학습 모델을 이용하여 감마값(γ) 및 알파값(α)의 특정 조합에 따른 가창 음성 데이터를 생성할 수 있다.
- [0104] 평가부(400)는 학습부(300)로부터 가창 음성 데이터를 수신할 수 있다.
- [0105] 평가부(400)는 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출할 수 있다.
- [0106] 평가부(400)는 상기 가창 음성 평가 점수를 기저장된 기준 평가 점수와 비교할 수 있다.
- [0107] 상기 기준 평가 점수는 최종 음원을 기준으로 청취자가 음원에 대한 이질감을 느끼지 않는 최소 점수를 의미하며, 사용자가 설정 가능한 값일 수 있다.
- [0108] 일 실시예에서, 평가부(400)는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 작은 경우, 학습 리스트 데이터 및 평가 리스트 데이터를 상기 입력부(100)의 입력으로 피드백할 수 있다.
- [0109] 이 경우, 입력부(100)는 가사 데이터, 미디 데이터, 음원 데이터, 학습 리스트 데이터, 및 평가 리스트 데이터를 파라미터 설정부(200)로 출력할 수 있다.
- [0110] 파라미터 설정부(200)는 상기 학습 리스트 데이터 및 상기 평가 리스트 데이터에 기초하여 상기 감마값(γ) 및 상기 알파값(α)을 재설정할 수 있다.
- [0111] 예를 들어, 재설정된 감마값(γ) 및 재설정된 알파값(α)은 학습부(300)의 BEGAN 기반 학습 모델을 통한 학습

리스트 데이터, 및 평가부(400)의 가창 음성 데이터의 음원 특성을 평가를 통한 평가 리스트 데이터가 반영된 최적화 파라미터일 수 있다.

- [0112] 일 실시예에서, 평가부(400)는 상기 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 상기 감마값(γ) 및 상기 알파값(α)을 최적화 파라미터로 저장할 수 있다.
- [0113] 또한, 가창 음성 평가 점수가 상기 기준 평가 점수보다 높은 경우, 평가부(400)는 가창 음성 데이터를 음원 추출부(500)에 출력할 수 있다.
- [0114] 음원 추출부(500)는 상기 가창 음성 데이터에 기초하여 최종 음원을 추출할 수 있다.
- [0115] 예를 들어, 음원 추출부(500)는 상기 평가부(400)가 최적화 파라미터로 저장하는 상기 감마값(γ) 및 상기 알파값(α)이 적용된 상기 가창 음성 데이터에 대하여 음원 튜닝을 수행함으로써 상기 최종 음원을 추출할 수 있다.
- [0116] 도 6은 감마값(γ) 및 알파값(α)의 임의의 조합을 나타내는 도표이고, 도 7은 도 6의 조합이 적용된 경우, 일 부의 스펙트로그램 및 파형을 나타내는 도면이다.
- [0117] 도 6을 참조하면, 감마값(γ)이 0.5, 0.75, 1.0 인 3가지 경우와 알파값(α)이 0, 0.5, 1.0인 3가지 경우를 각각 조합한 9가지 파라미터 조합이 예시되어 있다.
- [0118] 동일한 입력 데이터를 기준으로, 각각의 9가지 조합의 경우에 본 발명의 가창 음성 합성 장치(10)를 이용하여 가창 음성 데이터를 생성하고, 실제 시뮬레이션을 통해 가창 음성 데이터를 평가하였다.
- [0119] 평가 결과, 각각의 9가지 조합의 경우에 따른 가창 음성 평가 점수는 각각 상이한 결과를 나타냈다.
- [0120] 예를 들어, 도 7에서 보듯이, 감마값(γ) 및 알파값(α)이 각각 (0, 1.0)인 [4]번 조합의 경우, 음원의 인식이 불가능한 수준으로 낮은 품질을 나타내었다.
- [0121] 반면, 감마값(γ) 및 알파값(α)이 각각 (0.5, 0.5)인 [7]번 조합, 및 감마값(γ) 및 알파값(α)이 각각 (1.0, 1.0)인 [5]번 조합의 경우, 오디오 세그먼트의 스펙트로그램 및 음원 파형이 실제 사람이 부른 것과 같이 우수한 성능을 나타내었다.
- [0122] 이와 같이, 본 발명의 가창 음성 합성 장치(10)는 BEGAN 학습 모델을 이용하여 오토 인코더 아키텍처를 사용하여 손실 분포를 일치시키고, 판별자와 생성자 사이의 균형을 조정해주는 평형 개념(equilibrium concept)을 도입하여 생성 데이터의 다양성(diversity)과 품질(quality) 사이의 균형을 조정하며, 생성자 손실 함수(L_G)는 L1 손실의 가중치 역할을 수행하는 알파값(α)을 포함하여 생성자 손실의 과적합 문제를 최소화함으로써, 실제 사람이 부른 것 같이 자연스러운 고품질 가창 음성 음원을 생성할 수 있다.
- [0123] 도 8은 도 1의 가창 음성 합성 장치(10)를 이용한 가창 음성 합성 방법을 나타내는 순서도이다.
- [0124] 도 8을 참조하면, 본 발명의 가창 음성 합성 방법은 가사 데이터, 미디 데이터, 및 원곡 데이터를 포함하는 입력 데이터를 수신하는 단계(S100), 상기 입력 데이터에 기초하여 감마값(γ) 및 알파값(α)을 설정하는 단계(S200), 설정된 상기 감마값(γ) 및 상기 알파값(α)을 적용하여 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 단계(S300), 상기 가창 음성 데이터의 음원 특성을 평가하고, 가창 음성 평가 점수를 산출하는 단계(S400), 상기 가창 음성 평가 점수를 기저장된 기준 평가 점수와 비교하는 단계(S500), 가창 음성 평가 점수가 기저장된 기준 평가 점수보다 크거나 같은 경우, 현재 감마값(γ) 및 현재 알파값(α)을 최적화 파라미터로 저장하는 단계(S600), 및 상기 가창 음성 데이터에 기초하여 최종 음원을 추출하는 단계(S700)를 포함할 수 있다.
- [0125] 예를 들어, 상기 감마값(γ)은 상기 가창 음성 데이터의 품질과 다양성의 균형을 조절하는 하이퍼 파라미터이고, 상기 알파값(α)은 생성자 손실 함수에 포함된 L1 손실의 가중치일 수 있다.
- [0126] 일 실시예에서, 상기 원곡 데이터에 대한 타겟 가수의 가창 음성 데이터를 합성하는 단계는, BEGAN(Boundary Equilibrium Generative Adversarial Networks) 기반 학습 모델을 이용하여 상기 가창 음성 데이터를 합성할 수 있다.
- [0127] BEGAN 기반 학습 모델은 오토 인코더 아키텍처를 사용하여 손실 분포를 일치시킬 수 있다. 또한, BEGAN 기반 학습 모델은 판별자와 생성자 사이의 균형을 조정해주는 평형 개념(equilibrium concept)을 도입하여 생성 데이터의 다양성(diversity)과 품질(quality) 사이의 균형을 조정할 수 있다.

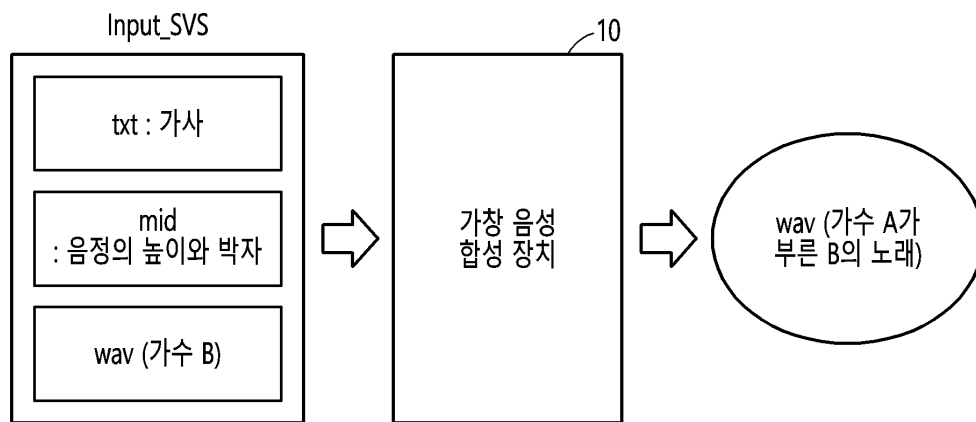
- [0128] 따라서, 본 발명의 가창 음성 합성 방법은 BEGAN 기반 학습 모델을 이용하여 생성자 손실의 과적합 문제를 최소화함으로써, 실제 사람이 부른 것 같이 자연스러운 고품질 가창 음성 음원을 생성할 수 있다.
- [0129] 다만, 이에 대해서는 상술한 바 있으므로, 그에 대한 중복되는 설명은 생략하기로 한다.
- [0130] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable 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 상기 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 설명된 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.
- [0131] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 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.
- [0132] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 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. 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

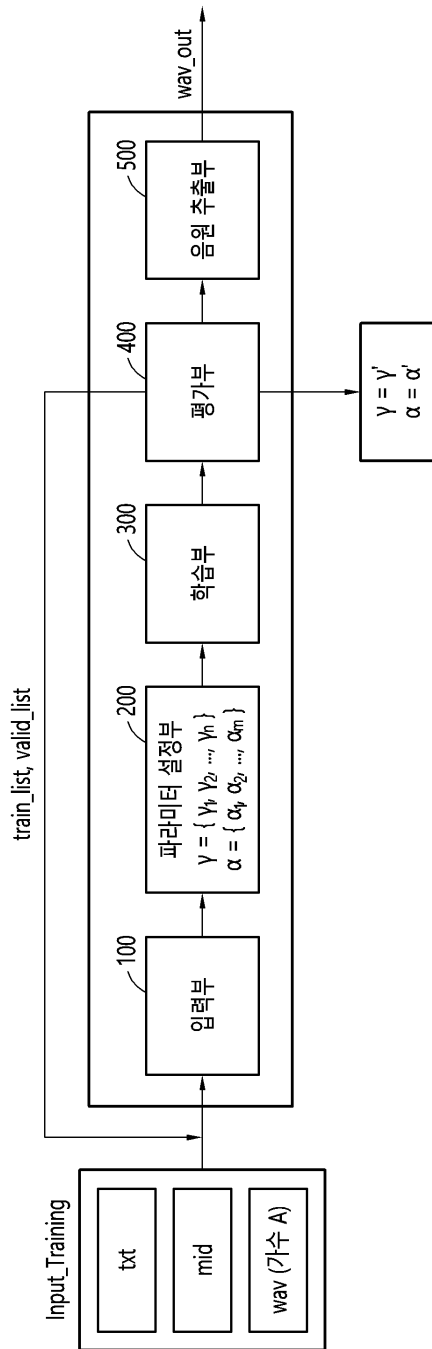
- [0133] 10: 가창 음성 합성 장치
- 100: 입력부
- 200: 파라미터 설정부
- 300: 학습부
- 400: 평가부
- 500: 음원 추출부

도면

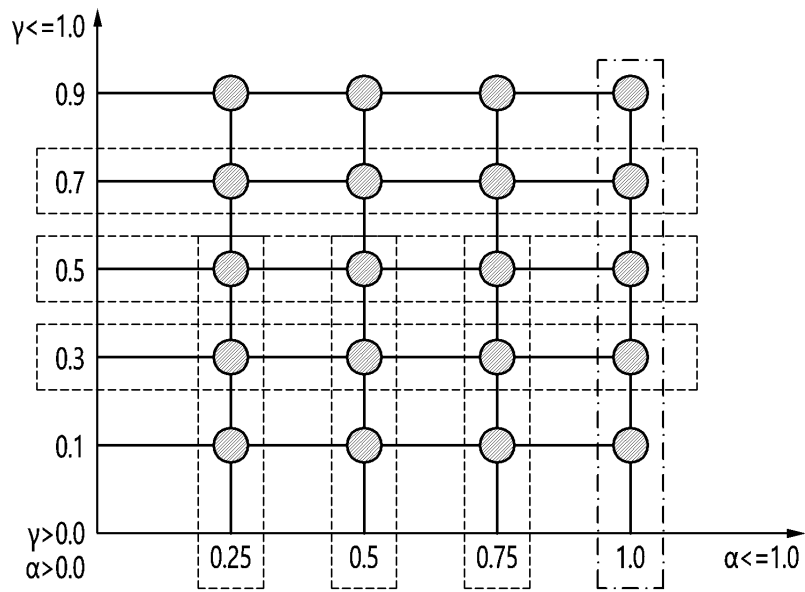
도면1



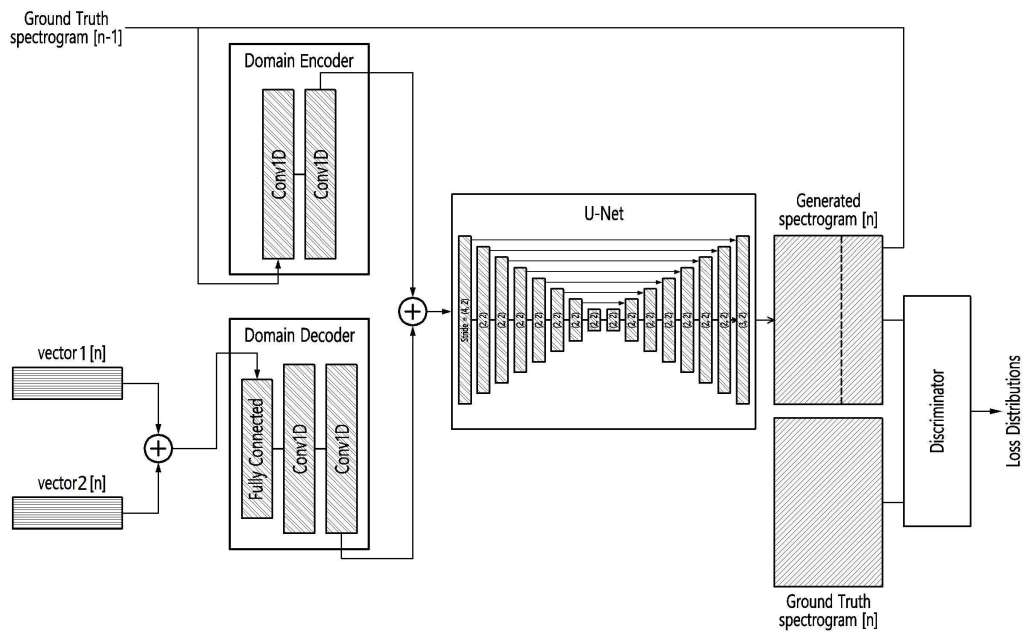
도면2



도면3



도면4



도면5

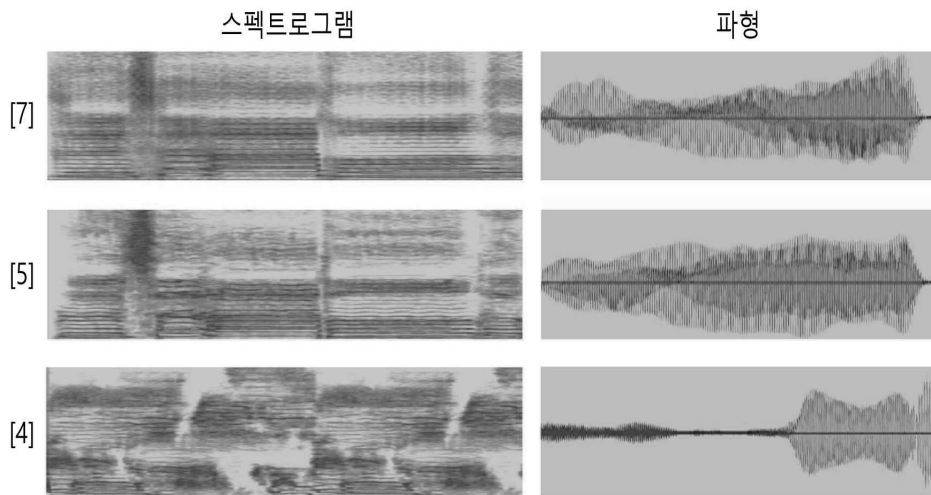
BEGAN-Sing objective function

$$\begin{cases} L_D = L(y) - k_t \cdot L(G(x)) & (1) \\ L_G = L(G(x)) + \alpha \cdot |y - G(x)| & (2) \\ k_{t+1} = k_t + \lambda_k (\gamma L(y) - L(G(x))) & (3) \end{cases}$$

도면6

$\alpha \backslash \gamma$	0.5	0.75	1.0
0.0	(0.0, 0.5) [2]	(0.0, 0.75) [6]	(0.0, 1.0) [4]
0.5	(0.5, 0.5) [7]	(0.5, 0.75) [1]	(0.5, 1.0) [8]
1.0	(1.0, 0.5) [9]	(1.0, 0.75) [3]	(1.0, 1.0) [5]

도면7



도면8

