(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0085188
(43) 공개일자 2025년06월12일

(51) 국제특허분류(Int. Cl.)
G10H 1/00 (2023.01) *G06F 16/432* (2019.01)
G06N 3/096 (2023.01)
(52) CPC특허분류
G10H 1/0008 (2023.01)
G06F 16/433 (2019.01)
(21) 출원번호 10-2023-0174176
(22) 출원일자 2023년12월05일
심사청구일자 2023년12월05일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
박인규
서울특별시 강남구 선릉로 221, 411동 2004호(도곡동, 도곡렉슬아파트)
매튜 마르첼루스
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(74) 대리인
양성보

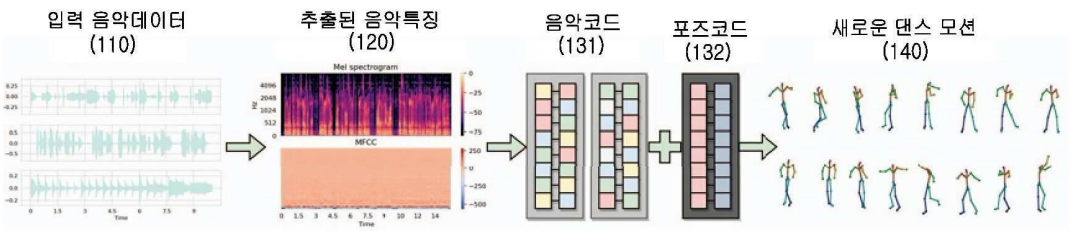
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **3차원 댄스 동작 생성을 위한 간결한 음악 표현 방법 및 시스템**

(57) 요 약

모션 생성을 위한 음악 표현 방법 및 시스템이 제시된다. 본 발명에서 제안하는 모션 생성을 위한 음악 표현 방법은 음악 특징 추출부를 통해 음악 데이터를 입력 받아 음악 특징을 추출하는 단계, 음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 단계, 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성하는 단계 및 모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G06F 16/4393 (2019.01)
G06N 3/0475 (2023.01)
G06N 3/096 (2023.01)
G06T 7/215 (2017.01)
G10H 2210/031 (2024.08)
G10H 2240/005 (2024.08)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193735
과제번호 2022-0-00981-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 방송통신산업기술개발사업
연구과제명 [Ezbaro] 전배경 정합 3D 객체 스트리밍 기술개발(2차년도)
기 여 율 9/10
과제수행기관명 인하대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711179347
과제번호 00155915
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신인재양성(교육운영수익)
연구과제명 [Ezbaro][정부+민간] 인공지능융합혁신인재양성(2차년도)
기 여 율 1/10
과제수행기관명 인하대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

음악 특징 추출부를 통해 음악 데이터를 입력 받아 음악 특징을 추출하는 단계;

음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 단계;

포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성하는 단계; 및

모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계

를 포함하는 모션 생성을 위한 음악 표현 방법.

청구항 2

제1항에 있어서,

상기 음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 단계는,

음악 인코더 블록을 통해 음악 특징의 시퀀스를 입력 받아 음악 특징의 시퀀스를 이산 변수를 사용하여 이산 음악 코드 시퀀스인 인코딩 특징 쌍의 시퀀스로 인코딩하고,

이산화 병목 블록을 통해 상기 인코딩 특징 쌍을 코드북 쌍 내에 가장 가까운 벡터로 양자화하며,

음악 디코더 블록을 통해 상기 가장 가까운 벡터를 디코딩하여 재구성 음악 시퀀스를 생성하며,

음악 인코더 블록과 코드북을 모션 생성을 위해 융합될 때까지 이산화 병목 블록의 손실과 함께 재구성 음악 시퀀스의 손실을 사용하여 상기 음악 코드 생성부를 훈련하는

모션 생성을 위한 음악 표현 방법.

청구항 3

제1항에 있어서,

상기 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성하는 단계는,

상기 음악 코드에 따른 포즈 코드를 생성하기 위해 상기 음악 코드를 다차원 특징으로 매핑하기 위한 한 쌍의 임베딩 공간을 추가하고, 상기 한 쌍의 임베딩 공간은 모든 다차원 특징에 적용되어 다차원 특징 쌍의 시퀀스를 생성하며, 상기 임베딩 공간은 상기 음악 코드에 따른 포즈 코드를 생성하기 위한 학습을 수행하고,

상기 임베딩 공간은 조건부 확률 샘플링 모듈과 통합하여 입력되는 모션과 음악 특징에 대한 분포를 근사화하고, 확률적 샘플링 측면을 보존하면서 역전파를 가능하게 하는 입력으로 노이즈를 샘플링하며, 이후 잔차 연결은 샘플링된 결과에 잠재 샘플링 결과를 추가하여 포즈 코드의 시퀀스를 생성하는

모션 생성을 위한 음악 표현 방법.

청구항 4

제3항에 있어서,

상기 모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계는,

상기 음악 코드, 과거 포즈 코드 및 상기 음악 코드에 따라 학습된 미래 포즈 코드 확률을 이용하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는

모션 생성을 위한 음악 표현 방법.

청구항 5

음악 데이터를 입력 받아 음악 특징을 추출하는 음악 특징 추출부;

상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 음악 코드 생성부;

상기 음악 코드에 따른 포즈 코드를 생성하는 포즈 코드 생성부; 및

상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 모션 생성부

를 포함하는 모션 생성을 위한 음악 표현 시스템.

청구항 6

제5항에 있어서,

상기 음악 코드 생성부는,

음악 인코더 블록을 통해 음악 특징의 시퀀스를 입력 받아 음악 특징의 시퀀스를 이산 변수를 사용하여 이산 음악 코드 시퀀스인 인코딩 특징 쌍의 시퀀스로 인코딩하고,

이산화 병목 블록을 통해 상기 인코딩 특징 쌍을 코드북 쌍 내에 가장 가까운 벡터로 양자화하며,

음악 디코더 블록을 통해 상기 가장 가까운 벡터를 디코딩하여 재구성 음악 시퀀스를 생성하며,

음악 인코더 블록과 코드북을 모션 생성을 위해 융합될 때까지 이산화 병목 블록의 손실과 함께 재구성 음악 시퀀스의 손실을 사용하여 상기 음악 코드 생성부를 훈련하는

모션 생성을 위한 음악 표현 시스템.

청구항 7

제5항에 있어서,

상기 포즈 코드 생성부는,

상기 음악 코드에 따른 포즈 코드를 생성하기 위해 상기 음악 코드를 다차원 특징으로 매핑하기 위한 한 쌍의 임베딩 공간을 추가하고, 상기 한 쌍의 임베딩 공간은 모든 다차원 특징에 적용되어 다차원 특징 쌍의 시퀀스를 생성하며, 상기 임베딩 공간은 상기 음악 코드에 따른 포즈 코드를 생성하기 위한 학습을 수행하고,

상기 임베딩 공간은 조건부 확률 샘플링 모듈과 통합하여 입력되는 모션과 음악 특징에 대한 분포를 근사화하고, 확률적 샘플링 측면을 보존하면서 역전파를 가능하게 하는 입력으로 노이즈를 샘플링하며, 이후 잔차 연결은 샘플링된 결과에 잠재 샘플링 결과를 추가하여 포즈 코드의 시퀀스를 생성하는

모션 생성을 위한 음악 표현 시스템.

청구항 8

제7항에 있어서,

상기 모션 생성부는,

상기 음악 코드, 과거 포즈 코드 및 상기 음악 코드에 따라 학습된 미래 포즈 코드 확률을 이용하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는

모션 생성을 위한 음악 표현 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 3차원 댄스 동작 생성을 위한 간결한 음악 표현 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 컴퓨터 그래픽의 최신 발전은 사람들이 AR/VR, 비디오 게임, 가상 세계, 소셜 네트워크 및 컴퓨터 제작 영화를 포함하여 컴퓨터 그래픽 보조 엔터테인먼트에 빠져들 수 있도록 한다. 이러한 서비스가 다양한 엔터테인먼트 선택 사항을 제공하지만 특히 3D 컴퓨터 그래픽에 기반한 서비스에 댄스를 통합하는 것은 어렵다. 그것은 종종 복잡한 인간 모션 캡처 시스템과 동작을 수행하는 전문 댄서를 필요로 한다. 미디어 소비율이 계속 증가하고 인터넷 트렌드가 계속 변화함에 따라 이러한 서비스에서 댄스 동작의 품질을 유지하는 것은 어려운 일이다.

[0003] 음악 조건부 3D 댄스 생성은 짧은 댄스 시퀀스와 음악 입력에서 댄스 모션을 생성하는 것을 포함하는 어려운 컴퓨터 비전 연구이다. 특히, 이 작업은 인간의 운동학적 제약으로 인해 복잡하며 예술적 창의성이 필요하다. 사실, 음악 작품에 수반되도록 고품질의 댄스 모션을 안무하는 것은 일반적으로 전문 안무가가 수행하는 학습되고 훈련된 기술이다. 종래기술에서는 심층 다중 모델을 사용하여 두 입력 데이터 형식(즉, 음악 데이터 및 조건화 댄스 모션)과 추정되는 후속 댄스 모션 사이의 상호 의존성을 학습하여 시퀀스 생성 작업으로 이 문제에 접근했다. 이러한 방법은 각각 섬세하게 설계된 심층 신경망을 사용하는 고유한 댄스 생성 방법을 제안한다. 그러나 종래기술의 공통적인 특성은 네트워크 설계를 통해 입력 음악 특징에 포함된 숨겨진 정보를 명시적으로 활용하지 못한다는 점이다.

선행기술문헌

비특허문헌

[0004] (비특허문헌 0001) [1] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Proc. Advances in Neural Information Processing Systems, pages 6306-6315, 2017.

(비특허문헌 0002) [2] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11040-11049, 2022.

발명의 내용

해결하려는 과제

[0005] 본 발명이 이루고자 하는 기술적 과제는 음악 특징에 대한 포괄적인 분석을 제시하고 음악 특징 내의 스케일 불균형 문제를 해결하기 위한 정규화 절차를 제안하는 3차원 댄스 동작 생성을 위한 간결한 음악 표현 방법 및 시스템을 제공하는데 있다. 또한, 기존 음악 특징 프로세서를 대체하는 음악 코드 생성부로 VQ-VAE를 응용한 M2C(Music-to-Codes) 네트워크를 제안한다. 본 발명의 실시예에 따른 접근 방식의 효과를 평가하기 위해 M2C를 SoTA 댄스 생성 방법을 수정한 확률적 Motion GPT(SM-GPT)와 결합하였다.

과제의 해결 수단

[0006] 일 측면에 있어서, 본 발명에서 제안하는 모션 생성을 위한 음악 표현 방법은 음악 특징 추출부를 통해 음악 데이터를 입력 받아 음악 특징을 추출하는 단계, 음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 단계, 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성하는 단계 및 모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계를 포함한다.

[0007] 상기 음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 단계는 음악 인코더 블록을 통해 음악 특징의 시퀀스를 입력 받아 음악 특징의 시퀀스를 이산 변수를 사용하여 이산 음악 코드 시퀀스인 인코딩 특징 쌍의 시퀀스로 인코딩하고, 이산화 병목 블록을 통해 상기 인코딩 특징 쌍을 코드북 쌍 내에 가장 가까운 벡터로 양자화하며, 음악 디코더 블록을 통해 상기 가장 가까운 벡터를 디코딩하여 재구성 음악 시퀀스를 생성하며, 음악 인코더 블록과 코드북을 모션 생성을 위해 융합될 때까지 이산화 병목 블록의 손실과 함께 재구성 음악 시퀀스의 손실을

사용하여 상기 음악 코드 생성부를 훈련한다.

[0008] 상기 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성하는 단계는 상기 음악 코드에 따른 포즈 코드를 생성하기 위해 상기 음악 코드를 다차원 특징으로 매핑하기 위한 한 쌍의 임베딩 공간을 추가하고, 상기 한 쌍의 임베딩 공간은 모든 다차원 특징에 적용되어 다차원 특징 쌍의 시퀀스를 생성하며, 상기 임베딩 공간은 상기 음악 코드에 따른 포즈 코드를 생성하기 위한 학습을 수행하고, 상기 임베딩 공간은 조건부 확률 샘플링 모듈과 통합하여 입력되는 모션과 음악 특징에 대한 분포를 근사화하고, 확률적 샘플링 측면을 보존하면서 역전파를 가능하게 하는 입력으로 노이즈를 샘플링하며, 이후 잔차 연결은 샘플링된 결과에 잠재 샘플링 결과를 추가하여 포즈 코드의 시퀀스를 생성한다.

[0009] 상기 모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계는 상기 음악 코드, 과거 포즈 코드 및 상기 음악 코드에 따라 학습된 미래 포즈 코드 확률을 이용하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성한다.

[0010] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 모션 생성을 위한 음악 표현 시스템은 음악 데이터를 입력 받아 음악 특징을 추출하는 음악 특징 추출부, 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 음악 코드 생성부, 상기 음악 코드에 따른 포즈 코드를 생성하는 포즈 코드 생성부 및 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 모션 생성부를 포함한다.

발명의 효과

[0011] 본 발명의 실시예들에 따른 3차원 댄스 동작 생성을 위한 간결한 음악 표현 방법 및 시스템을 통해 음악 특징에 대한 포괄적인 분석을 제시하고 음악 특징 내의 스케일 불균형 문제를 해결하기 위한 정규화 절차를 제안한다. 음악 코드 생성부로 VQ-VAE를 응용한 M2C(Music-to-Codes) 네트워크를 제안하여 기존 음악 특징 프로세서를 대체하고, M2C를 SoTA 댄스 생성 방법을 수정한 확률적 Motion GPT(SM-GPT)와 결합하여 제안하는 접근 방식의 효과를 평가한다.

도면의 간단한 설명

[0012] 도 1은 본 발명의 일 실시예에 따른 3차원 댄스 동작 생성을 위한 간결한 음악 표현 과정을 설명하기 위한 개략도이다.

도 2는 본 발명의 일 실시예에 따른 모션 생성을 위한 음악 표현 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 모션 생성을 위한 음악 표현 시스템의 구성을 나타내는 도면이다.

도 4는 본 발명의 일 실시예에 따른 M2C 및 SM-GPT를 포함하는 네트워크 파이프라인을 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시예에 따른 MFCC 계수의 평균값을 나타내는 그래프이다.

도 6은 본 발명의 일 실시예에 따른 M2C 네트워크 아키텍처를 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0013] 음악과 동기화된 3D 댄스 모션을 생성하는 것은 음악 리듬과 인체 움직임 사이의 복잡한 상호 작용을 모델링하는 것이 수반되기 때문에 어려운 작업이다. 대부분의 기존 접근 방식은 댄스 모션 생성에 중요한 역할을 하는 음악 특징 처리 단계의 중요성을 간과하는 경우가 많다. 본 발명에서는 이산 변수를 사용하여 음악 특징을 더욱 잠재적으로 표현하는 음악 코드를 제안한다. 본 발명의 실시예에 따르면, 음악 특징에 대한 포괄적인 분석을 제시하고 음악 특징 내의 스케일 불균형 문제를 해결하기 위한 다른 정규화 절차를 제안한다. 또한, 기존 음악 특징 프로세서를 대체하는 음악 코드 추출기로 VQ-VAE를 응용한 M2C(Music-to-Codes) 네트워크를 제안한다.

[0014] 본 발명의 실시예에 따르면, 음악 특징의 근본적인 측면에 대한 종합적인 분석을 제시하고, 특히 음악 조건부 댄스 생성에서의 적용을 강조한다. 본 발명의 실시예에 따른 새로운 음악 코드를 공식화하기 위해 M2C 네트워크를 제안한다. 본 발명의 실시예에 따르면, 음악 코드와의 호환성을 보장하기 위해 SoTA 댄스 예측기를 수정하고, 생성된 댄스 동작의 다양성을 높이기 위해 확률적 샘플링 모듈을 통합한다.

[0015] 본 발명의 실시예에 따른 접근 방식의 효과를 평가하기 위해 M2C를 최근 SoTA 댄스 생성 방법을 수정한 확률적 Motion GPT(SM-GPT)와 결합한다. 본 발명의 실시예에 따른 광범위한 평가 및 어블레이션(ablation) 연구는 제안

하는 댄스 생성 파이프라인(M2C 및 SMGPT 사용)이 모든 평가 지표에서 댄스 생성 결과를 질적, 양적으로 크게 향상시킨다는 것을 확인하였다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0017] 도 1은 본 발명의 일 실시예에 따른 3차원 댄스 동작 생성을 위한 간결한 음악 표현 과정을 설명하기 위한 개략도이다.

[0018] 본 발명에서는 음악 입력 내에 필수 특징을 캡처하여 심층 댄스 생성 모델에 도움이 되도록 설계된 간결한 음악 표현을 제안한다. 본 발명의 실시예에 따르면, 입력되는 음악 데이터(110)를 이용하여 음악 특징(120)을 추출하고, 추출된 음악 특징을 이용하여 표현을 음악 코드(131)로 레이블링한다. 이러한 음악 코드(131)는 이산 변수를 이용한 이산 잠재 코드 쌍이다. 또한, 본 발명에서는 음악 특징(120)을 음악 코드(131)로 대체할 필요성과 이점에 관한 연구 결과를 제시하고, 일반적으로 사용되는 추출된 음악 특징 MFCC(Mel-Frequency Cepstral Coefficients)의 대규모 스케일 불균형을 보여주는 증거를 제공한다.

[0019] 본 발명의 실시예에 따르면, 음악 특징 시퀀스(VQ-VAE 및 3D 포즈 VQ-VAE을 이용함)가 주어진 음악 코드 공식화를 위한 M2C(Music-to-Codes) 네트워크를 제안한다. 본 발명의 실시예에 따르면, M2C를 이산 오토인코더로 설계하고 댄스 생성을 위한 음악 코드를 공식화하는 데 사용하기 전에 훈련한다. 이후, M2C의 특징 추출기를 댄스 동작을 생성하기 위해 수정된 SoTA 심층 댄스 생성 네트워크인 Li의 Motion GPT를 통해 생성된 포즈 코드(132)와 결합한다. M2C와 SM-GPT 네트워크의 전체 결합 구조는 도 4에 도시하였다. 수정에는 음악 코드와의 호환성을 보장하고 다양성을 높이기 위한 확률적 샘플링 모듈을 통합하는 것이 포함된다. 간단한 설계에도 불구하고 실험 결과는 M2C가 생성된 댄스 동작 품질을 향상시키는 통찰력 있는 이산 음악 코드를 성공적으로 공식화한다는 것을 보여준다.

[0020] 본 발명의 실시예에 따른 모션 생성과 음악 조건부 댄스 생성에 있어서, 사람의 모션을 합성하기 위한 모션 그래프 기반 접근법을 채택한 초기 연구와 함께 수년 동안 사람의 모션을 합성하는 광범위한 연구가 수행되었다. 이러한 접근 방식은 전체 기록된 모션 시퀀스를 하위 시퀀스로 분할하여 만든 모션 그래프를 결합하여 사람의 모션을 합성한다. 그러나 음악 조건부 댄스 생성은 그럴듯한 댄스 시퀀스를 생성할 뿐만 아니라 주어진 음악 작품과 일치시키기 위해 댄스 모션과 음악적 속성 사이의 교차 모드 이해가 필요하다. 최근에는 음악적 속성과 해당 댄스 동작 사이의 상관 관계를 학습하기 위해 CNN, RNN, GAN 및 트랜스포머 네트워크와 같은 댄스 생성에 학습 기반 방법을 사용했다.

[0021] 종래기술에서는 인과 주의 마스크와 달리 전체 주의 마스크가 있는 트랜스포머 네트워크를 사용하여 댄스 생성 방법을 제안하고 음악 및 댄스 장르 주석과 함께 3D 및 2D 관절 주석이 있는 대규모 댄스 모션 데이터셋 AIST++를 생성했다. 종래기술에서는 일반적인 애니메이션 기법에서 영감을 받아 각 비트의 키 프레임을 먼저 예측한 다음 TCB 스플라인을 사용하여 키 프레임 간 보간하여 댄스 모션을 표현하도록 설계된 또 다른 트랜스포머 기반 댄스 생성 네트워크를 제안했다. 최근 연구 중 하나로 사람의 포즈의 강력한 이산 표현을 먼저 학습한 다음 각 관절을 직접 조작하는 대신 이산 표현을 예측하는 2단계 댄스 생성 네트워크가 제안되었다. 또 다른 종래기술에서는 음악 특징을 추출하기 위해 이전에 수작업으로 만든 오디오 특징 추출 방법을 활용하면서 확산 기반 접근 방식을 사용하여 음악에서 편집 가능한 댄스 생성을 제안했다.

[0022] 앞서 언급한 방법 중 많은 방법이 네트워크 입력의 구성 요소로 음악 속성을 얻었다. 제안하는 음악 코드 아이디어와는 달리, 종래기술에서는 추출한 음악 특징을 네트워크에 직접 공급하여 해당 음악 특징과 댄스 동작 사이의 적절한 매핑을 학습했다. 반면, 또 다른 종래기술에서는 강력한 입력 음악 특징 추출 방법을 활용하여 본 발명과 일부 유사성을 공유한다. 그러나 강력한 오디오 특징 추출기를 사용하거나 이산 특징을 사용하지는 않는다.

[0023] 본 발명의 실시예에 따른 이산 잠재 표현에 있어서, 잠재 특징의 한 가지 변형은 인코딩 네트워크(예를 들어, 변형 자동 인코더)를 사용하여 생성된 연속 잠재 변수이다. 이산 잠재 표현은 언어의 이산 특성(즉, 단어 토큰)을 고려할 때 언어 처리 태스크에 더 일반적으로 사용된다. 이와 함께, 일부 최근 컴퓨터 비전 작업은 이산 잠재 공간을 활용하는 방법을 성공적으로 설계했다. 종래기술에서는 악명 높은 GAN 훈련 어려움 없이 현대 GAN에 필적하는 고품질 이미지를 생성하기 위해 다중 스케일 계층 VQ-VAE를 설계했다. 또 다른 종래기술에서는 이산 VAE를 이미지 자동 인코더로 사용하는 제로 샷 텍스트 대 이미지 방법을 제안했다. 또한 이산 VAE가 트랜스포머 네트워크의 컨텍스트 크기를 192배 감소시키는 것으로 나타났다. 마지막으로 또 다른 종래기술에서는 모션-VQ 네트워크 내에서 VQ-VAE 설계를 활용하여 여러 사람 포즈를 두 개의 포즈 코드로 인코딩 했다.

[0024] 본 발명의 실시예에 따른 VQ-VAE를 통한 음악 특징 모델링 에 있어서, 오디오 데이터에 대해 VQ-VAE를 활용한 몇 가지 연구가 있다. 종래기술에서는 새로운 음악 생성 태스크를 위한 다양한 시간 해상도를 모델링하기 위해 여러 개의 개별 VQ-VAE 모델(Hierarchical VQ-VAE에서 영감을 받아)을 활용했다. 또 다른 종래기술에서는 VQ-VAE를 기반으로 생성 모델을 제안했으며, 이 모델은 큰 소리를 분리하고 주어진 음색 분포를 양자화한다.

[0025] 본 발명의 실시예에 따른 M2C 네트워크는 음악 특징을 양자화하여 이산 표현을 추출하기 위해 VQ-VAE를 기반으로 설계되었기 때문에 이러한 이전 방법과 유사성을 공유한다. 종래기술과 달리 M2C 이면의 동기는 양자화된 인코딩 특징을 활용하는 것이 아니라 음악 코드 시퀀스를 공식화하는 것이다. 이는 SM-GPT 내에서 훈련 가능한 코드북을 사용하여 각 음악 코드에서 보다 적절한 특징 표현을 학습할 때 분명해진다.

[0027] 도 2는 본 발명의 일 실시예에 따른 모션 생성을 위한 음악 표현 방법을 설명하기 위한 흐름도이다.

[0028] 본 발명의 실시예에 따른 모션 생성을 위한 음악 표현 방법은 음악 특징 추출부를 통해 음악 데이터를 입력 받아 음악 특징을 추출하는 단계(210), 음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타내는 단계(220), 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성하는 단계(230) 및 모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계(240)를 포함한다.

[0029] 단계(210)에서, 음악 특징 추출부를 통해 음악 데이터를 입력 받아 음악 특징을 추출한다.

[0030] 단계(220)에서, 음악 코드 생성부를 통해 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타낸다.

[0031] 본 발명의 실시예에 따르면, 음악 인코더 블록을 통해 음악 특징의 시퀀스를 입력 받아 음악 특징의 시퀀스를 이산 변수를 사용하여 이산 음악 코드 시퀀스인 인코딩 특징 쌍의 시퀀스로 인코딩한다.

[0032] 본 발명의 실시예에 따르면, 이산화 병목 블록을 통해 상기 인코딩 특징 쌍을 코드북 쌍 내에 가장 가까운 벡터로 양자화한다.

[0033] 본 발명의 실시예에 따르면, 음악 디코더 블록을 통해 상기 가장 가까운 벡터를 디코딩하여 재구성 음악 시퀀스를 생성한다.

[0034] 본 발명의 실시예에 따르면, 음악 인코더 블록과 코드북을 모션 생성을 위해 융합될 때까지 이산화 병목 블록의 손실과 함께 재구성 음악 시퀀스의 손실을 사용하여 상기 음악 코드 생성부를 훈련한다.

[0035] 단계(230)에서, 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성한다.

[0036] 본 발명의 실시예에 따르면, 상기 음악 코드에 따른 포즈 코드를 생성하기 위해 상기 음악 코드를 다차원 특징으로 매핑하기 위한 한 쌍의 임베딩 공간을 추가하고, 상기 한 쌍의 임베딩 공간은 모든 다차원 특징에 적용되어 다차원 특징 쌍의 시퀀스를 생성하며, 상기 임베딩 공간은 상기 음악 코드에 따른 포즈 코드를 생성하기 위한 학습을 수행한다.

[0037] 본 발명의 실시예에 따르면, 상기 임베딩 공간은 조건부 확률 샘플링 모듈과 통합하여 입력되는 모션과 음악 특징에 대한 분포를 근사화하고, 확률적 샘플링 측면을 보존하면서 역전파를 가능하게 하는 입력으로 노이즈를 샘플링하며, 이후 잔차 연결은 샘플링된 결과에 잠재 샘플링 결과를 추가하여 포즈 코드의 시퀀스를 생성한다.

[0038] 단계(240)에서, 모션 생성부를 통해 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성한다.

[0039] 본 발명의 실시예에 따르면, 상기 음악 코드, 과거 포즈 코드 및 상기 음악 코드에 따라 학습된 미래 포즈 코드 확률을 이용하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성한다.

[0041] 도 3은 본 발명의 일 실시예에 따른 모션 생성을 위한 음악 표현 시스템의 구성을 나타내는 도면이다.

[0042] 본 실시예에 따른 모션 생성을 위한 음악 표현 시스템(300)은 프로세서(310), 버스(320), 네트워크 인터페이스(330), 메모리(340) 및 데이터베이스(350)를 포함할 수 있다. 메모리(340)는 운영체제(341) 및 모션 생성을 위한 음악 표현 루틴(342)을 포함할 수 있다. 프로세서(310)는 음악 특징 추출부(311), 음악 코드 생성부(312),

포즈 코드 생성부(313) 및 모션 생성부(314)를 포함할 수 있다. 다른 실시예들에서 모션 생성을 위한 음악 표현 시스템(300)은 도 3의 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, 모션 생성을 위한 음악 표현 시스템(300)은 디스플레이나 트랜시버(transceiver)와 같은 다른 구성요소들을 포함할 수도 있다.

[0043] 메모리(340)는 컴퓨터에서 판독 가능한 기록 매체로서, RAM(random access memory), ROM(read only memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록장치(permanent mass storage device)를 포함할 수 있다. 또한, 메모리(340)에는 운영체제(341)와 모션 생성을 위한 음악 표현 루틴(342)을 위한 프로그램 코드가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 드라이브 메커니즘(drive mechanism, 미도시)을 이용하여 메모리(340)와는 별도의 컴퓨터에서 판독 가능한 기록 매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록 매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록 매체(미도시)를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록 매체가 아닌 네트워크 인터페이스(330)를 통해 메모리(340)에 로딩될 수도 있다.

[0044] 버스(320)는 모션 생성을 위한 음악 표현 시스템(300)의 구성요소들간의 통신 및 데이터 전송을 가능하게 할 수 있다. 버스(320)는 고속 시리얼 버스(high-speed serial bus), 병렬 버스(parallel bus), SAN(Storage Area Network) 및/또는 다른 적절한 통신 기술을 이용하여 구성될 수 있다.

[0045] 네트워크 인터페이스(330)는 모션 생성을 위한 음악 표현 시스템(300)을 컴퓨터 네트워크에 연결하기 위한 컴퓨터 하드웨어 구성요소일 수 있다. 네트워크 인터페이스(330)는 모션 생성을 위한 음악 표현 시스템(300)을 무선 또는 유선 커넥션을 통해 컴퓨터 네트워크에 연결시킬 수 있다.

[0046] 데이터베이스(350)는 모션 생성을 위한 음악 표현을 위해 필요한 모든 정보를 저장 및 유지하는 역할을 할 수 있다. 도 3에서는 모션 생성을 위한 음악 표현 시스템(300)의 내부에 데이터베이스(350)를 구축하여 포함하는 것으로 도시하고 있으나, 이에 한정되는 것은 아니며 시스템 구현 방식이나 환경 등에 따라 생략될 수 있고 혹은 전체 또는 일부의 데이터베이스가 별개의 다른 시스템 상에 구축된 외부 데이터베이스로서 존재하는 것 또한 가능하다.

[0047] 프로세서(310)는 기본적인 산술, 로직 및 모션 생성을 위한 음악 표현 시스템(300)의 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(340) 또는 네트워크 인터페이스(330)에 의해, 그리고 버스(320)를 통해 프로세서(310)로 제공될 수 있다. 프로세서(310)는 음악 특징 추출부(311), 음악 코드 생성부(312), 포즈 코드 생성부(313) 및 모션 생성부(314)를 위한 프로그램 코드를 실행하도록 구성될 수 있다. 이러한 프로그램 코드는 메모리(340)와 같은 기록 장치에 저장될 수 있다.

[0048] 음악 특징 추출부(311), 음악 코드 생성부(312), 포즈 코드 생성부(313) 및 모션 생성부(314)는 도 2의 단계들(210~240)을 수행하기 위해 구성될 수 있다.

[0049] 모션 생성을 위한 음악 표현 시스템(300)은 음악 특징 추출부(311), 음악 코드 생성부(312), 포즈 코드 생성부(313) 및 모션 생성부(314)를 포함할 수 있다.

[0050] 본 발명의 실시예에 따른 음악 특징 추출부(311)는 음악 데이터를 입력 받아 음악 특징을 추출한다.

[0051] 본 발명의 실시예에 따른 음악 코드 생성부(312)는 상기 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 상기 추출된 음악 특징을 음악 코드로 나타낸다.

[0052] 본 발명의 실시예에 따르면, 음악 인코더 블록을 통해 음악 특징의 시퀀스를 입력 받아 음악 특징의 시퀀스를 이산 변수를 사용하여 이산 음악 코드 시퀀스인 인코딩 특징 쌍의 시퀀스로 인코딩한다.

[0053] 본 발명의 실시예에 따르면, 이산화 병목 블록을 통해 상기 인코딩 특징 쌍을 코드북 쌍 내에 가장 가까운 벡터로 양자화한다.

[0054] 본 발명의 실시예에 따르면, 음악 디코더 블록을 통해 상기 가장 가까운 벡터를 디코딩하여 재구성 음악 시퀀스를 생성한다.

[0055] 본 발명의 실시예에 따르면, 음악 인코더 블록과 코드북을 모션 생성을 위해 융합될 때까지 이산화 병목 블록의 손실과 함께 재구성 음악 시퀀스의 손실을 사용하여 상기 음악 코드 생성부를 훈련한다.

[0056] 본 발명의 실시예에 따른 포즈 코드 생성부(313)는 포즈 코드 생성부를 통해 상기 음악 코드에 따른 포즈 코드를 생성한다.

[0057] 본 발명의 실시예에 따른 포즈 코드 생성부(313)는 상기 음악 코드에 따른 포즈 코드를 생성하기 위해 상기 음악 코드를 다차원 특징으로 매핑하기 위한 한 쌍의 임베딩 공간을 추가하고, 상기 한 쌍의 임베딩 공간은 모든 다차원 특징에 적용되어 다차원 특징 쌍의 시퀀스를 생성하며, 상기 임베딩 공간은 상기 음악 코드에 따른 포즈 코드를 생성하기 위한 학습을 수행한다.

[0058] 본 발명의 실시예에 따른 포즈 코드 생성부(313)는 상기 임베딩 공간은 조건부 확률 샘플링 모듈과 통합하여 입력되는 모션과 음악 특징에 대한 분포를 근사화하고, 확률적 샘플링 측면을 보존하면서 역전파를 가능하게 하는 입력으로 노이즈를 샘플링하며, 이후 잔차 연결은 샘플링된 결과에 잠재 샘플링 결과를 추가하여 포즈 코드의 시퀀스를 생성한다.

[0059] 본 발명의 실시예에 따른 모션 생성부(314)는 상기 음악 코드와 포즈 코드를 결합하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성한다.

[0060] 본 발명의 실시예에 따른 모션 생성부(314)는 상기 음악 코드, 과거 포즈 코드 및 상기 음악 코드에 따라 학습된 미래 포즈 코드 확률을 이용하여 상기 입력된 음악 데이터에 따른 댄스 모션을 생성한다.

[0062] 도 4는 본 발명의 일 실시예에 따른 M2C 및 SM-GPT를 포함하는 네트워크 파이프라인을 설명하기 위한 도면이다.

[0063] 본 발명의 실시예에 따르면, 일련의 음악 특징을 사용하여 음악 코드의 이산 매핑을 학습하기 위해 M2C 네트워크를 제안한다. 실제로 M2C 네트워크는 모든 음악 조건 댄스 생성 방법에서 모든 음악 특징 인코더 모듈(예를 들어, FC 계층 및 1D 컨볼루션 계층)을 대체할 수 있는 대체 모듈 역할을 한다. 이는 특정 요구 사항 없이 기존 음악 특징 인코더를 대체하기 위한 다목적 옵션인 모든 음악 특징 조합에 적합할 수 있기 때문에 가능하다. 그러나 댄스 생성을 위해 M2C 네트워크를 사용하려면 댄스 생성을 위해 특별히 설계된 다른 네트워크가 필요하다. 이를 위해 본 발명에서는 M2C 네트워크를 SoTA 댄스 생성 네트워크인 Motion GPT의 수정된 버전인 SM-GPT와 결합한다. 도 4는 M2C(410)와 SM-GPT(420)를 모두 포함하는 본 발명의 실시예에 따른 평가 네트워크의 전체 파이프라인을 보여준다.

[0064] 본 발명의 실시예에 따르면, 댄스 생성을 위한 핵심 구성 요소로서 음악 특징 분석을 위한 연구 결과를 설명한다. 먼저, 이전 댄스 생성 작업이 음악 특징을 추출하기 위해 댄스 생성 태스크를 위한 일반적인 음악 특징 추출 도구인 공공 오디오 처리 도구 상자를 이용한다. 각 댄스 생성 방법은 특정 음악 특징 조합을 음악 입력으로 고안한다.

[0065] MFCC는 음색 질감 특징이나 스펙트럼 특징을 추출하여 음성 처리 태스크에 널리 사용된다. 추출된 음색 특성은 음악 처리 태스크에도 특히 유용하다. 그러나 MFCC는 리듬 관련 특징을 구성하지 않는다. 따라서 대부분의 댄스 생성 방법은 다른 음악 특징과 쌍을 이룬다.

[0067] 도 5는 본 발명의 일 실시예에 따른 MFCC 계수의 평균값을 나타내는 그래프이다.

[0068] 도 5에서 음악 특징값 분포의 특성이 딥 러닝 네트워크에 직접 공급되기에 부적합함을 보여준다. 예를 들어, MFCC의 첫 번째 계수는 오프셋 값으로 나머지 MFCC들과 평균 및 표준 편차가 상당히 다르다. 값 스케일링 불균형은 다른 계수들에 비해 처음 5개의 MFCC에 더 큰 의미를 부여한다. 이는 값 정규화가 중요할 뿐만 아니라 종종 방법의 성능에 결정적인 요소가 되므로 딥 러닝 원리와 일치하지 않는다.

[0069] 이를 완화하기 위해 본 발명에서는 특징 벡터의 인덱스에 따라 음악 특징을 정규화한다(이러한 정규화 방법을 새로운 norm이라고 함). 이는 전체 벡터 차원에서 특징을 정규화하는 표준 정규화 기법에서 벗어난다는 점에 주목한다. 본 발명의 실시예에 따른 음악 특징 정규화 방법을 다음과 같이 공식화한다:

[0070]
$$\mathbf{m}_{norm} = \left\{ \left\{ \frac{\mathbf{m}_{d,n} - \underset{i \in N}{Mean}(\mathbf{m}_{d,i})}{\underset{i \in N}{Std}(\mathbf{m}_{d,i})} \right\}_{d=0}^{D} \right\}_{n=0}^{N} \quad (1)$$

[0071] 여기서 m은 음악 특징, D는 음악 특징 벡터의 전체 차원, N은 훈련 또는 테스트 데이터셋 내에서 음악 샘플과 음악 시퀀스 양의 합, Mean(m_d) 및 Std(m_d)는 특정 인덱스 d 내에서 각 음악 특징의 평균 및 표준 편차를 나타낸다. 본 발명에서 제안하는 정규화 함수는 균일한 값 분포를 달성하여 정규화 후에도 특징 벡터 샘플 전반에 걸

쳐 유지되는 표준 정규화 기법에서 관찰된 스케일 문제를 해결한다.

[0073] 도 6은 본 발명의 일 실시예에 따른 M2C 네트워크 아키텍처를 나타내는 도면이다.

[0074] 본 발명의 일 실시예에 따른 VQVAE를 응용한 M2C(Music-to-Codes Network)는, 음악 특징 시퀀스(610)를 입력받아 음악 특징 시퀀스 $\mathbf{m} = \langle m_t \rangle_{t=1}^{T}$를 인코딩 특징 쌍 $\mathbf{h} = \langle h_t^1, h_t^2 \rangle_{t=1}^{T}$의 시퀀스로 인코딩하는 음악 인코더 블록 E(x)(620), 코드북 쌍 $C = \{\{e_k\}_{k=1}^{K}\}_{n=1}^{N=2}$ 내에 가장 가까운 벡터 $\mathbf{e}_k = \langle e_{k_t}^1, e_{k_t}^2 \rangle_{t=1}^{T}$로 $\mathbf{h} \mapsto \mathbf{e}_k$를 양자화하는 이산화 병목 블록(Discretization Bottleneck)(630), e를 디코딩하여 재구성 음악 시퀀스 $\hat{\mathbf{m}}$(650)를 생성하는 음악 디코더 블록 D(x)(640)를 포함한다.

[0075] 본 발명의 실시예에 따른 M2C는 시간적 도메인을 유지하면서 m의 공간적 도메인만을 인코딩함으로써, 한 쌍의 이산 음악 코드 시퀀스 $\mathbf{k} = \langle k_t^1, k_t^2 \in [K] \rangle_{t=1}^{T}$을 생성하고, 여기서 K는 어휘 키 크기를 나타낸다. 각 코드북 $C_{n=1}^{N=2}$은 h^n 중 하나와 상호 작용하여 e^n과 k^n 쌍을 만든다. 먼저 M2C 인코더 E(x)와 코드북 C를 댄스 생성에 활용하기 위해 융합될 때까지 M2C를 훈련한다.

[0076] 본 발명의 실시예에 따르면, 이산화 병목 손실(즉, $L_{codebook}$ 및 L_{commit})과 함께 표준 재구성 손실 Lrec.를 사용하여 M2C를 훈련한다. 재구성 손실 $L_{rec.}(\hat{\mathbf{m}}, \mathbf{m})$은 재구성된 음악 특징 시퀀스 $\hat{\mathbf{m}}$과 입력 음악 특징 시퀀스 m 사이의 거리를 계산한다. VQ-VAE에 따라 이산화 병목 손실을 다음과 같이 정의한다.

[0077] $$L_{codebook} = \frac{1}{T}\sum_{t} ||sg[\mathbf{h}_t] - \mathbf{e}_{k_t}||_2^2 \quad (2)$$

[0078] 키 k와 h 내의 코드북 특징 e_k 사이의 거리를 줄이려 한다. sg[.]는 정지 경사 프로세스를 나타내고 손실은 시간 도메인 T를 향해 평균화된다.

[0079] $$L_{commit} = \frac{1}{T}\sum_{t} ||\mathbf{h}_t - sg[\mathbf{e}_{k_t}]||_2^2 \quad (3)$$

[0080] 인코딩 특징 h가 특정 임베딩 공간에 커밋되도록 유도하여 값 변동을 제한하려 한다. 위의 모든 손실을 결합하면 M2C 훈련 목표가 생성된다:

[0081] $$L_{M2C} = L_{rec.}(\hat{\mathbf{m}}, \mathbf{m}) + \alpha \times (L_{commit} + L_{codebook}) + \beta \quad (4)$$

[0082] 여기서 α는 $L_{codebook}$ 및 L_{commit} 의 하이퍼 파라미터이고 β는 정규화 값이다.

[0083] 다시 도 4를 참조하면, 본 발명의 실시예에 따른 M2C의 음악 코드를 활용하는 댄스 예측 모듈로 Motion GPT(425)를 활용한다. 이러한 이유로 음악 코드 k와의 호환성을 보장하고 수정된 버전을 SM-GPT(420)로 레이블을 지정하기 위해 Motion GPT에 대한 몇 가지 수정 사항을 추가했다. 먼저 음악 코드를 다차원 특징 $\mathbf{k} \mapsto \tilde{\mathbf{e}}_k$.로 매핑하기 위해 한 쌍의 임베딩 공간 $\tilde{C} = \{\{e_k\}_{k=1}^{K}\}_{n=1}^{N=2}$을 추가한다(421). 임베딩 공간 쌍 $\tilde{C}$는 모든 $\sum_{n}^{2}\sum_{t}^{T}\{\mathbf{k}\}_t^n$에 적용되어 다차원 특징 쌍 $\tilde{\mathbf{e}} = \langle \tilde{e}_t^1, \tilde{e}_t^2 \rangle_{t=1}^{T}$의 시퀀스를 생성한다(422). 이 임베딩 공간은 입력 음악 특징 처리를 위한 Motion GPT의 조밀한 계층을 대체한다. Motion GPT 내에 학습 가능한 전용 임베딩 공간이 있으면 m과 $\tilde{\mathbf{e}}_k$ 간의 단절이 발생하여 추출된 음악 코드를 기반으로 각 입력 음악 샘플에 대한 보다 적절한 표현을 학습할 수 있는 기회가 만들어진다.

[0084] 또한, 댄스 생성 다양성을 향상시키기 위해 공간 $\tilde{\mathbf{C}}$를 임베딩한 후 직접 조건부 확률 샘플링 모듈을 통합한다. 일반적인 CVAE와 유사하게 본 발명에서는 $q_\theta(\mathbf{z}|\tilde{\mathbf{e}}, \mathbf{x}, \mathbf{y})$을 훈련하기 위한 posterior 네트워크(423)와

$p_\theta(\mathbf{z}|\tilde{\mathbf{e}},\mathbf{x})$을 테스트하기 위한 prior 네트워크(424)의 두 가지 주요 경로를 가진 잔차 확률 샘플링 모듈(잔차 샘플링으로 레이블 지정)을 설계한다. posterior 네트워크 $q_\theta(\mathbf{z}|\tilde{\mathbf{e}},\mathbf{x},\mathbf{y})$는 타겟 댄스 모션 y의 도움을 받아 입력 댄스 모션 x와 음악 특징 $\tilde{\mathbf{e}}$에 대한 posterior 분포를 근사화한다. 한편, prior 네트워크 $p_\theta(\mathbf{z}|\tilde{\mathbf{e}},\mathbf{x})$는 테스트 시간 동안 타겟 정보 y가 부족하기 때문에 y가 없는 x와 $\tilde{\mathbf{e}}$에 대한 prior 분포를 근사화한다. posterior 네트워크와 prior 네트워크는 모두 적층된 다층 퍼셉트론(MLP)으로 구성된 특징 추출에 동일한 인코더 Esamp(x)를 사용한다. 본 발명에서는 재파라미터화 트릭을 활용하고 확률적 샘플링 측면을 보존하면서 역전파를 가능하게 하는 입력으로 노이즈를 샘플링한다. 이후 잔차 연결은 Esamp(x)의 결과에 잠재 샘플링 결과 z를 추가하여 포즈 코드 시퀀스와 함께 모션 GPT에 조건부 입력 c를 생성한다.

[0085] 음악 조건화 댄스 생성에서의 문제는 조건화 댄스 모션 $\mathbf{x} = \langle x_t \rangle_{t=1}^{T}$과 조건화 음악 특징 $\mathbf{m} = \langle m_t \rangle_{t=1}^{T}$이 주어진 새로운 댄스 시퀀스 $\hat{\mathbf{y}} = \langle y_t \rangle_{t=1}^{T}$을 생성하는 것이다.

[0086] 도 4를 다시 참조하면, 앞서 언급했듯이 본 발명의 실시예에서는 최근 SoTA, Motion GPT의 수정된 버전인 SM-GPT를 댄스 생성 네트워크로 활용한다. 또한, 새로운 댄스 생성 프로세스에 대해 음악 코드 $\mathbf{k} = \langle k_t^1, k_t^2 \rangle_{t=1}^{T}$를 추출하기 위해 사전 훈련된 M2C를 활용한다. 이 훈련 단계에서 SM-GPT만 훈련하고 M2C 인코더와 이산화 병목을 활용하여 음악 코드를 추출한다. 도 4는 M2C를 사용한 새로운 댄스 생성을 위한 전체 네트워크 아키텍처를 보여준다.

[0087] 새로운 댄스 동작을 생성하기 위해 먼저 정규화된 음악 입력 m_{norm}에서 M2C(410)를 사용하여 음악 코드 k(414)를 추출한다. 이는 M2C 인코더 E(x)(412)를 사용하여 입력 음악 특징 시퀀스 m_t(411)를 $\mathbf{h}_t^n$로 인코딩한 후, 이산화 병목 블록(413)을 통해 인코딩된 특징 $\mathbf{h}_t^n$를 각각의 임베딩 공간 $C_{n=1}^{N=2}$ 내에 있는 양자화된 특징 $\mathbf{e}_t^n$으로 양자화하고 음악 코드 $\mathbf{k}_t^n$(414)를 획득한다. 앞서 설명된 바와 같이, M2C는 시간적 도메인은 그대로 둔 채 공간적 도메인만 인코딩하여, 음악 코드 추출 과정 내내 입력 음악 특징 벡터 내의 시간 단계 t를 영향 없이 유지한다.

[0088] SM-GPT(420)는 추출된 음악 코드 $\mathbf{k}_t^n$를 코드북 쌍 $\tilde{C}$를 사용하여 다차원 특징으로 변환한다. 이 코드북 쌍 $\tilde{C}$는 음악 코드 $\mathbf{k}_t^n$를 해석하기 위해 SM-GPT 내에서 훈련 가능한 임베딩 공간이다(M2C의 코드북 쌍 C와 혼동되지 않음). 각 음악 코드에 대한 구별되고 전용 임베딩 공간을 갖는 것은 비선형성을 도입하고 네트워크가 보다 적절한 표현을 학습할 수 있도록 한다. 이후 잔차 샘플링 모듈(훈련 중 posterior 네트워크 또는 테스트 중 prior 네트워크 사용)에서 조건부 입력 c_t를 얻는다. 모션 GPT는 3D 포즈 VQ-VAE의 조건부 입력 c_t와 포즈 코드 시퀀스를 사용하여 새로운 댄스 모션을 생성한다.

[0089] 본 발명의 실시예에 따른 Motion GPT의 훈련 목표를 채택하여 새로운 댄스 생성 네트워크를 훈련시키는 동시에 잔차 샘플링 모듈을 훈련시키는 가변 학습 목표 SGVB(Stochastic Gradient Variational Bayes)를 추가한다. z가 대각선 공분산 행렬이 있는 다변량 가우스 분포를 따른다고 가정하면 훈련 목표를 증거 하한(Evidence Lower Bound; ELBO)으로 쓸 수 있다

[0090] $$L_{ELBO} = L_{CE} - KL(q_\theta(\mathbf{z}|\tilde{\mathbf{e}},\mathbf{x},\mathbf{y})||p_\theta(\mathbf{z}|\tilde{\mathbf{e}},\mathbf{x})) \quad (5)$$

[0091] 여기서 L_{CE}는 본 발명의 재구성 손실에 이어 posterior 분포와 prior 분포 사이의 KL(Kullback-Libler) 발산이다. Motion GPT를 따라 L_{CE}를 다음과 같이 정의한다.

[0092] $$L_{CE} = \frac{1}{T}\sum_{t=1}^{T}\sum_{h=u,l} CrossEntropy(\mathbf{a}_t^h, p_{t+1}^h) \quad (6)$$

[0093] 여기서 a_t는 과거 포즈 코드 x_t 및 음악 코드 $\mathbf{k}_t^n$가 주어진 미래 포즈 코드 확률을 나타내고, p_{t+1}은 GT 미래 포

즈 코드, u 및 l은 각각 상체 또는 하체 포즈 코드를 나타낸다. 본 발명에서는 KL 가중치를 점진적으로 증가시킴으로써 네트워크가 정확성 있게 훈련할 수 있도록 하는 KL 비용 어닐링(cost annealing)을 통합한다.

[0095] 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, 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 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.

[0096] 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 장치에 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.

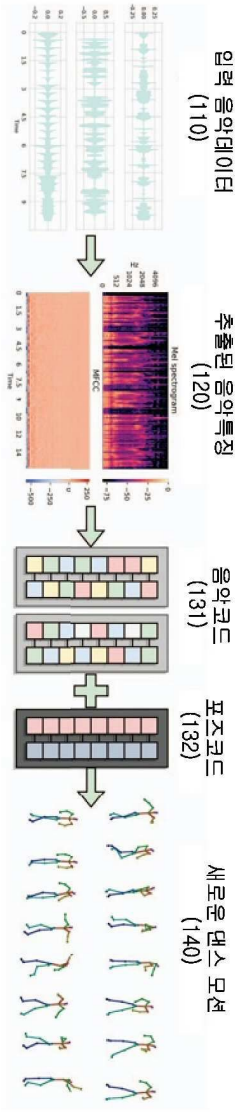
[0097] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic 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.

[0098] 이상과 같이 실시예들이 비록 한정된 실시예와 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.

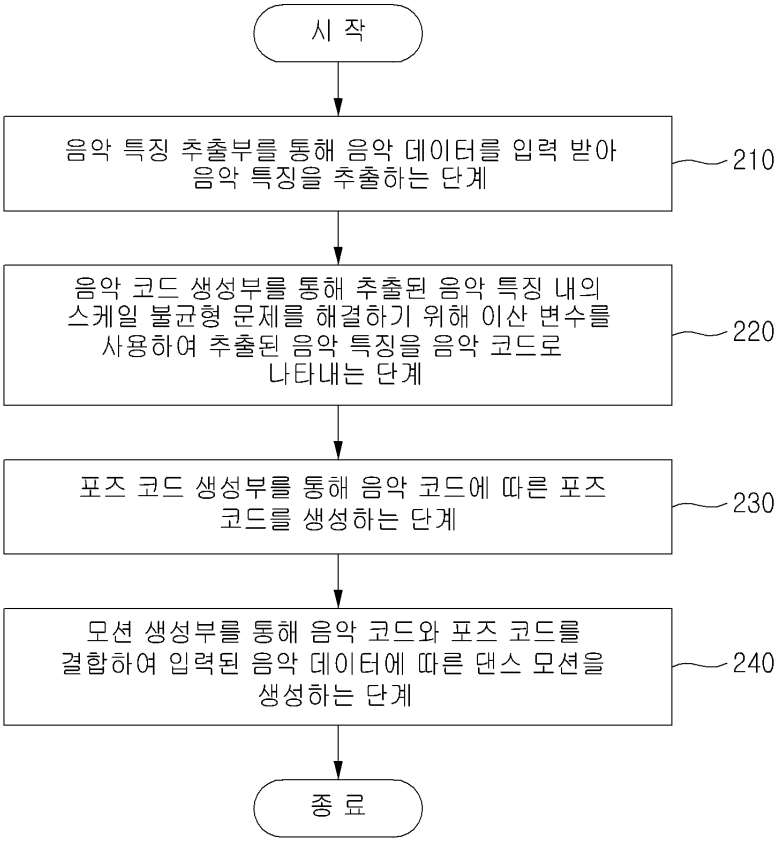
[0099] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

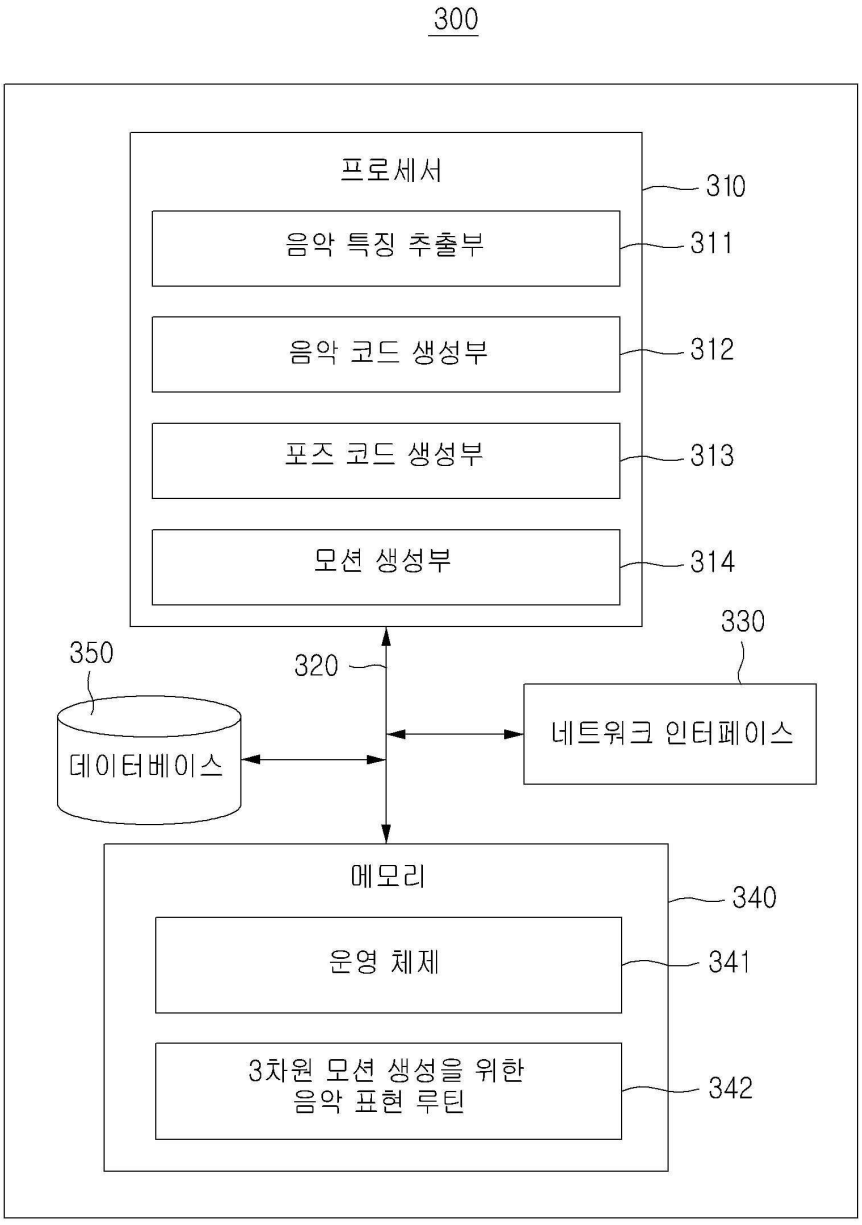
도면1



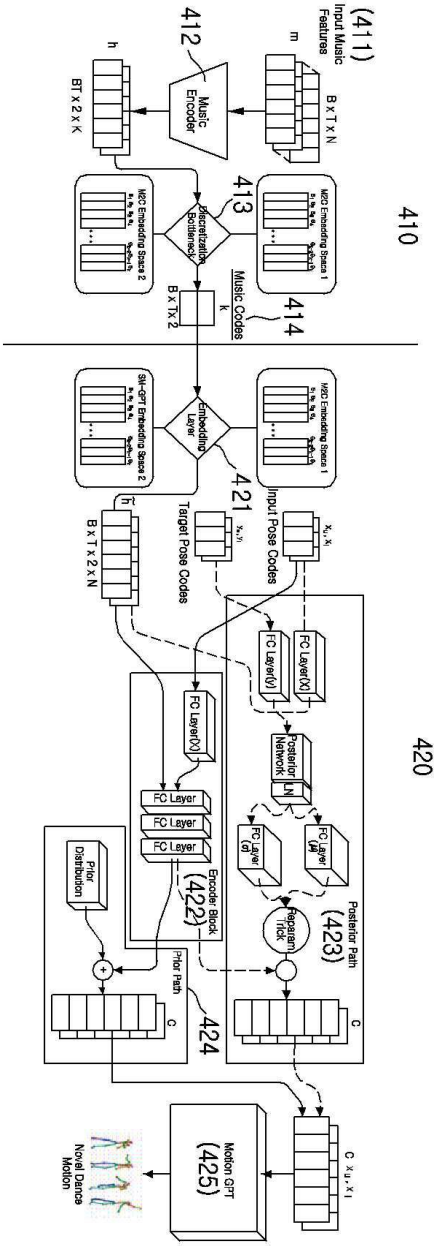
도면2


시 작
음악 특징 추출부를 통해 음악 데이터를 입력 받아 음악 특징을 추출하는 단계
210
음악 코드 생성부를 통해 추출된 음악 특징 내의 스케일 불균형 문제를 해결하기 위해 이산 변수를 사용하여 추출된 음악 특징을 음악 코드로 나타내는 단계
220
포즈 코드 생성부를 통해 음악 코드에 따른 포즈 코드를 생성하는 단계
230
모션 생성부를 통해 음악 코드와 포즈 코드를 결합하여 입력된 음악 데이터에 따른 댄스 모션을 생성하는 단계
240
종 료

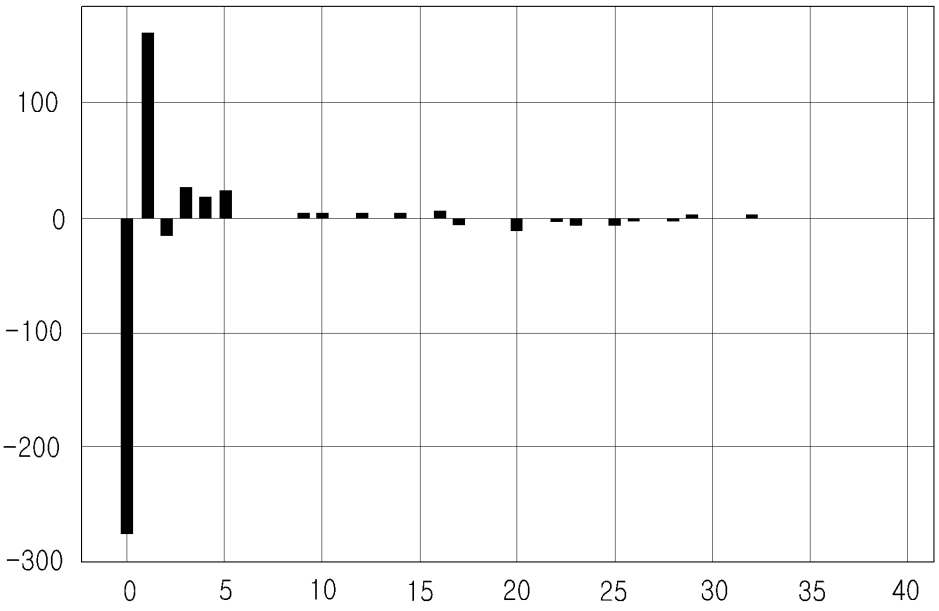
도면3


300
프로세서
310
음악 특징 추출부
311
음악 코드 생성부
312
포즈 코드 생성부
313
모션 생성부
314
330
350
320
네트워크 인터페이스
데이터베이스
메모리
340
운영 체제
341
3차원 모션 생성을 위한
음악 표현 루틴
342

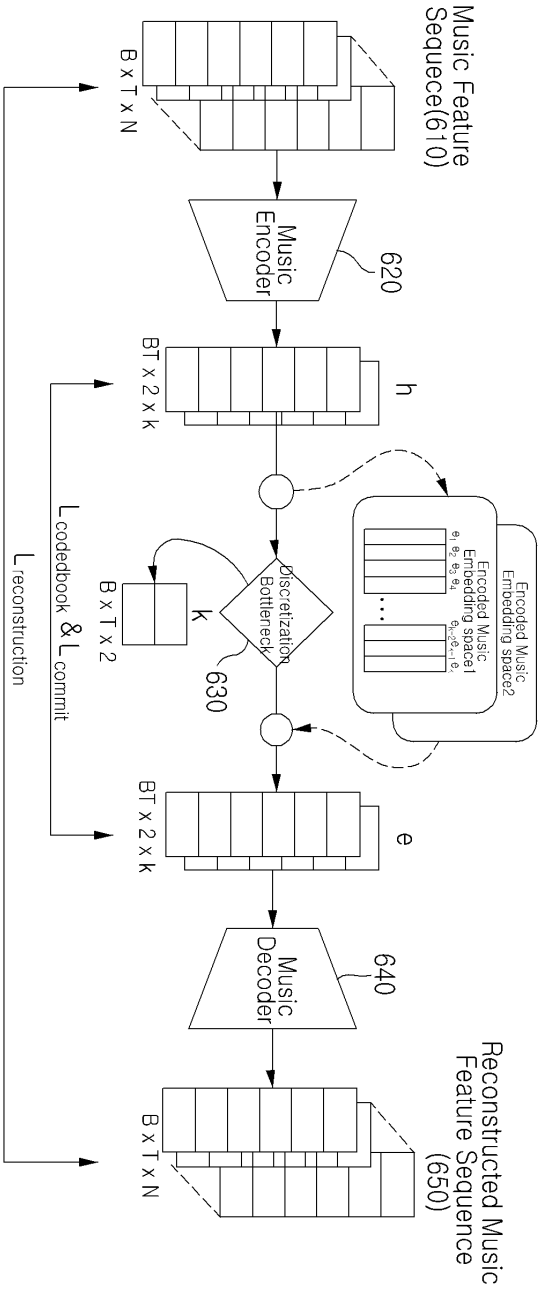
도면4


(411)
Input Music Features
412
Music Encoder
413
Music Codes
414
410
Input Pose Codes
Target Pose Codes
421
FC Layer(X)
FC Layer(V)
Posterior Network
FC Layer
Prior Distribution
Encoder Block
(422)
Prior Path
424
Reparam Trick
(423)
Posterior Path
420
Motion GPT
(425)
Novel Dance Motion

도면5



도면6


Music Feature Sequece(610)
B x T x N
Music Encoder
620
BT x 2 x k
h
Encoded Music Embedding space2
Encoded Music Embedding space1
Discretization Bottleneck
B x T x 2
k
630
L codebook & L commit
L reconstruction
BT x 2 x k
e
Music Decoder
640
Reconstructed Music Feature Sequence (650)
B x T x N