(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년07월15일
(11) 등록번호 10-2277205
(24) 등록일자 2021년07월08일

(51) 국제특허분류(Int. Cl.)
G10L 13/033 (2013.01) *G10L 15/02* (2006.01)
G10L 25/63 (2013.01)
(52) CPC특허분류
G10L 13/033 (2013.01)
G10L 15/02 (2013.01)
(21) 출원번호 10-2020-0033038
(22) 출원일자 2020년03월18일
심사청구일자 2020년03월18일
(56) 선행기술조사문헌
KR1020160049804 A*
KR102057927 B1*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
휴멜로 주식회사
서울특별시 강남구 역삼로 169 ,2층(명우빌딩) (역삼동)
(72) 발명자
엘기르, 모하메드 이메드 모하메드 카멜
이집트, 12591, 기자 가버너레이트, 식스 오브 악토버, 포스 디스트릭트, 스트리트 5, 1212
박중배
서울특별시 강남구 역삼로18길 12, 401호(역삼동)
(74) 대리인
특허법인가산

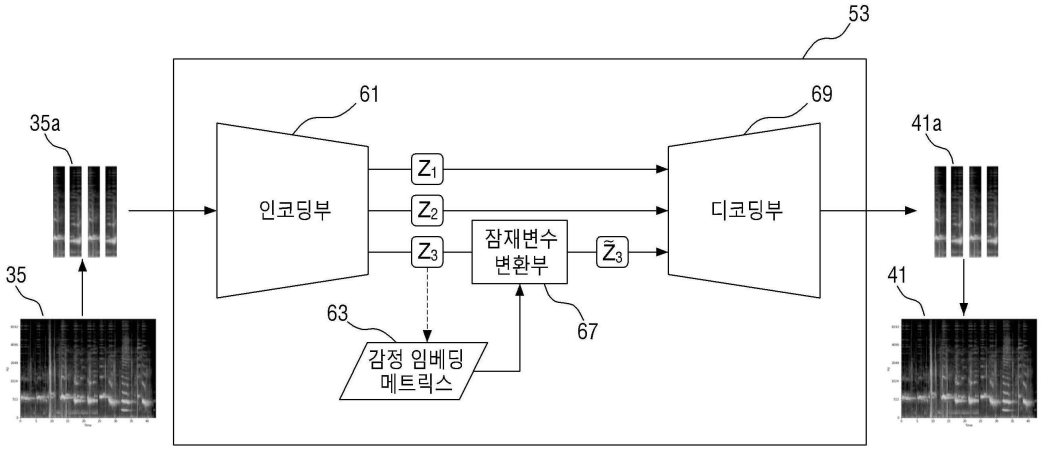
전체 청구항 수 : 총 16 항 심사관 : 정성윤

(54) 발명의 명칭 **오디오 변환 장치 및 방법**

(57) 요 약

텍스트 생성 모델을 이용하여 입력 텍스트로부터 출력 텍스트를 생성하는 방법이 제공된다. 본 발명의 일 실시예에 따른 텍스트 생성 방법은, 입력 텍스트에 포함된 적어도 하나의 단어를 가리키는 데이터를 상기 텍스트 생성 모델에 입력하여 잠재 변수(latent variable)를 획득하는 단계와, 상기 잠재 변수를 이용하여 상기 출력 텍스트에 포함될 제1 타깃 단어가 속하는 타깃 군집을 예측하는 단계와, 상기 타깃 군집 및 상기 잠재 변수를 이용하여 상기 제1 타깃 단어를 예측하는 단계를 포함한다.

대 표 도



(52) CPC특허분류
G10L 25/30 (2013.01)
G10L 25/63 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	0000190019
과제번호	CY190019
부처명	산업통상자원부
과제관리(전문)기관명	서울특별시 서울산업진흥원
연구사업명	서울시 산학연 협력사업(2019년 인공지능 기술사업화 지원사업)
연구과제명	인공지능 음성 변환/합성 기술 기반 프리미엄 콘텐츠 생성 기술 개발
기 여 율	1/1
과제수행기관명	휴멜로 주식회사
연구기간	2019.10.01 ~ 2020.09.30

명 세 서

청구범위

청구항 1

컴퓨팅 장치가 신경망 기반 오디오 변환 모델을 이용하여 입력 오디오 시퀀스의 특징을 변환하는 방법으로서,

상기 입력 오디오 시퀀스를 나타내는 입력 오디오 데이터를 상기 오디오 변환 모델에 입력하는 단계;

상기 오디오 변환 모델에 의해, 상기 입력 오디오 시퀀스의 시퀀스 레벨 특징을 나타내는 제1 잠재 변수(latent variable)를 획득하는 단계;

상기 오디오 변환 모델에 의해, 상기 입력 오디오 시퀀스를 구성하는 복수의 세그먼트들 각각의 특징들을 나타내는 제2 잠재 변수를 획득하는 단계;

상기 제1 잠재 변수를 조정하는 단계; 및

상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 이용하여, 출력 오디오 시퀀스를 나타내는 출력 오디오 데이터를 획득하는 단계

를 포함하되,

상기 제1 잠재 변수는, 상기 입력 오디오 시퀀스에 표현된 화자의 감정을 적어도 부분적으로 표현하는 제3 잠재 변수를 포함하고,

상기 제1 잠재 변수를 조정하는 단계는, 제1 감정에 대응되는 임베딩 벡터를 제2 감정에 대응되는 임베딩 벡터로 변환하는 감정 변환 함수를 상기 제3 잠재 변수에 적용하는 단계를 포함하는,

오디오 변환 방법.

청구항 2

제1항에 있어서,

상기 제1 잠재 변수는,

상기 입력 오디오 시퀀스에 표현된 상기 화자의 감정 이외의 음향적 특징을 적어도 부분적으로 표현하는 제4 잠재 변수를 더 포함하고,

상기 제2 잠재 변수는 ,

상기 입력 오디오 시퀀스에 포함된 언어적 컨텐츠(linguistic content)를 적어도 부분적으로 표현하는 것이며,

상기 제1 잠재 변수를 조정하는 단계는,

상기 제4 잠재 변수를 그대로 유지한 채 상기 제3 잠재 변수를 조정하는 단계를 포함하는,

오디오 변환 방법.

청구항 3

삭제

청구항 4

삭제

청구항 5

제1항에 있어서,

상기 감정 변환 함수를 상기 제3 잠재 변수에 적용하는 단계는,

상기 제1 감정에 대응되는 임베딩 벡터 및 상기 제2 감정에 대응되는 임베딩 벡터를 직교화(orthogonalization)하는 단계를 더 포함하는,

오디오 변환 방법.

청구항 6

제1항에 있어서,

상기 출력 오디오 시퀀스는,

상기 입력 오디오 시퀀스의 언어적 컨텐츠 및 상기 입력 오디오 시퀀스의 화자의 음성의 특징을 유지한 채, 상기 입력 오디오 시퀀스에 표현된 상기 화자의 감정이 변환된 오디오인,

오디오 변환 방법.

청구항 7

제2항에 있어서,

상기 제1 잠재 변수를 획득하는 단계는,

상기 오디오 변환 모델의 인코딩부에 의해, 상기 입력 오디오 데이터를 이용하여 상기 제3 잠재 변수 예측 분포를 획득하는 단계; 및

상기 제3 잠재 변수 예측 분포로부터 상기 제3 잠재 변수를 샘플링하는 단계

를 포함하는,

오디오 변환 방법.

청구항 8

제7항에 있어서,

상기 제3 잠재 변수 예측 분포는 정규 분포를 따르도록 학습된 것인,

오디오 변환 방법.

청구항 9

제7항에 있어서,

상기 제1 잠재 변수를 획득하는 단계는,

상기 인코딩부에 의해, 상기 입력 오디오 데이터를 이용하여 상기 제4 잠재 변수 예측 분포를 획득하는 단계; 및

상기 제4 잠재 변수 예측 분포로부터 상기 제4 잠재 변수를 샘플링하는 단계

를 더 포함하는,

오디오 변환 방법.

청구항 10

제1항에 있어서,

상기 제2 잠재 변수를 획득하는 단계는,

상기 입력 오디오 데이터 및 상기 제1 잠재 변수를 이용하여, 상기 제2 잠재 변수를 획득하는 단계를 포함하는,

오디오 변환 방법.

청구항 11

제1항에 있어서,

상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 이용하여, 출력 오디오 데이터를 획득하는 단계는,

상기 오디오 변환 모델의 디코딩부에 의해, 상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 결합(concatenate)한 값을 이용하여, 상기 출력 오디오 데이터를 예측하기 위한 분포를 획득하는 단계; 및

상기 출력 오디오 데이터를 예측하기 위한 상기 분포로부터 상기 출력 오디오 데이터를 샘플링하는 단계

를 포함하는,

오디오 변환 방법.

청구항 12

제11항에 있어서,

상기 출력 오디오 데이터를 보코더에 입력하여 출력 오디오 시퀀스를 합성하는 단계

를 더 포함하는,

오디오 변환 방법.

청구항 13

오디오를 변환하는 장치로서,

인코딩부, 잠재 변수 변환부, 및 디코딩부를 포함하는 오디오 변환 모델; 및

출력 오디오 데이터로부터 출력 오디오 시퀀스를 합성하는 오디오 합성부

를 포함하되,

상기 인코딩부는,

입력 오디오 시퀀스를 전처리한 입력 오디오 데이터로부터, 상기 입력 오디오 시퀀스의 시퀀스 레벨 특징을 나타내는 제1 잠재 변수를 인코딩하여, 상기 제1 잠재 변수를 상기 잠재 변수 변환부에 제공하고,

상기 입력 오디오 데이터 및 상기 제1 잠재 변수로부터, 상기 입력 오디오 시퀀스를 구성하는 복수의 세그먼트들 각각의 특징들을 나타내는 제2 잠재 변수를 인코딩하여, 상기 제2 잠재 변수를 상기 디코딩부에 제공하며,

상기 잠재 변수 변환부는,

상기 제1 잠재 변수를 조정하고, 조정된 제1 잠재 변수를 상기 디코딩부에 제공하고,

상기 디코딩부는,

상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 이용하여, 상기 출력 오디오 데이터를 제공하되,

상기 제1 잠재 변수는, 상기 입력 오디오 시퀀스에 표현된 화자의 감정을 적어도 부분적으로 표현하는 제3 잠재 변수를 포함하고,

상기 잠재 변수 변환부는, 제1 감정에 대응되는 임베딩 벡터를 제2 감정에 대응되는 임베딩 벡터로 변환하는 감정 변환 함수를 상기 제3 잠재 변수에 적용하여 상기 제1 잠재 변수를 조정하는,

오디오 변환 장치.

청구항 14

제13항에 있어서,

상기 인코딩부는,

상기 입력 오디오 데이터로부터 상기 제1 잠재 변수의 평균을 인코딩하고, 상기 제1 잠재 변수의 평균에 기초하여 결정되는 예측 분포로부터 상기 제1 잠재 변수를 샘플링하며,

상기 입력 오디오 데이터 및 상기 제1 잠재 변수로부터 상기 제2 잠재 변수의 평균을 인코딩하고, 상기 제2 잠

재 변수의 평균에 기초하여 결정되는 예측 분포로부터 상기 제2 잠재 변수를 샘플링하는,

오디오 변환 장치.

청구항 15

제13항에 있어서,

상기 제1 잠재 변수는,

상기 입력 오디오 시퀀스에 표현된 화자의 감정 이외의 음향적 특징을 적어도 부분적으로 표현하는 제4 잠재 변수를 더 포함하고,

상기 제2 잠재 변수는,

상기 입력 오디오 시퀀스에 포함된 언어적 컨텐츠를 적어도 부분적으로 표현하는 것이며,

상기 잠재 변수 변환부는, 상기 제4 잠재 변수를 그대로 유지한 채 상기 제3 잠재 변수를 조정하는,

오디오 변환 장치.

청구항 16

삭제

청구항 17

삭제

청구항 18

제13항에 있어서,

상기 디코딩부는,

상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 결합(concatenate)한 값을 이용하여, 상기 출력 오디오 데이터 예측 분포를 획득하고, 상기 출력 오디오 데이터 예측 분포로부터 상기 출력 오디오 데이터를 샘플링하는,

오디오 변환 장치.

청구항 19

컴퓨팅 장치가 신경망 기반 오디오 변환 모델을 트레이닝 하는 방법으로서,

복수의 오디오 시퀀스 및 상기 복수의 오디오 시퀀스 각각에 나타난 화자의 감정을 가리키는 레이블을 포함하는 학습 데이터 세트를 획득하는 단계; 및

상기 학습 데이터 세트를 이용하여 상기 오디오 변환 모델의 파라미터를 업데이트하는 단계

를 포함하되,

상기 오디오 변환 모델은,

입력 오디오 시퀀스를 인코딩하여 상기 입력 오디오 시퀀스에 나타난 화자의 감정을 나타내는 제1 잠재 변수(latent variable) 및 상기 입력 오디오 시퀀스를 구성하는 복수의 세그먼트들 각각의 특징들을 나타내는 제2 잠재 변수를 제공하고,

제1 감정에 대응되는 임베딩 벡터를 제2 감정에 대응되는 임베딩 벡터로 변환하는 감정 변환 함수를 상기 제1 잠재 변수에 적용하여, 상기 제1 잠재 변수를 조정하며,

상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 디코딩하여 상기 화자의 감정이 변환된 출력 오디오 시퀀스를 제공하는 것이며,

상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는,

제1 입력 오디오 시퀀스를 상기 오디오 변환 모델에 입력할 경우 상기 제1 입력 오디오 데이터와 동일한 오디오

시퀀스가 출력될 확률이 증가하도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 포함하고,

상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는,

상기 제1 잠재 변수 및 상기 제2 잠재 변수의 분포가 정규 분포에 가까워지도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 포함하는,

오디오 변환 모델 트레이닝 방법.

청구항 20

제19항에 있어서,

상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는,

각각 서로 다른 감정을 가리키는 레이블을 가지는 복수의 오디오 시퀀스들을 인코딩하여 획득되는 상기 제1 잠재 변수들의 거리가 서로 멀어지도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 더 포함하는,

오디오 변환 모델 트레이닝 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 오디오 변환 장치 및 방법에 관한 것이다. 보다 자세하게는, 신경망 기반의 오디오 변환 모델을 이용한 음성의 변환에 있어서, 화자의 목소리 등의 일부 특징들은 그대로 유지한 채, 음성에 표현된 화자의 감정을 변환하는 장치 및 방법과, 그 오디오 변환 모델을 구축하는 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 인공 신경망을 이용한 머신 러닝 기술은, 음성 합성(speech synthesis) 및 음성 변환(voice conversion) 등 다양한 기술 분야에 활용되고 있다. 음성 합성 기술은, 예컨대 입력된 텍스트로부터 사람이 말하는 소리와 유사한 소리를 합성해 내는 기술이다. 음성 변환 기술은, 예를 들어 발화의 스타일 중 일부를 인위적으로 변환하는 기술이다. 음성 변환 기술은, 오디오 북, 엔터테인먼트, 외국어 교육, 및 외화 더빙 등 다양한 분야에서 활용될 수 있다.

[0003] 종래의 음성 변환 기술은 음성 오디오에 담긴 화자의 메시지, 즉 언어적 컨텐츠(linguistic content)를 그대로 전달하면서 화자의 목소리를 변환하는 것이 대부분이다. 예를 들어, 특정 남성 화자(speaker)의 목소리가 담긴 오디오를 특정 여성 화자의 목소리로 변환하는 것이 음성 변환의 대표적인 예이다.

[0004] 한편, 음성 오디오에 담긴 화자의 메시지와 화자의 목소리를 그대로 유지한 채로, 음성의 다른 특징들, 예컨대 박자, 억양, 운율, 음성에 표현된 감정 등을 변환하려는 시도들도 이루어지고 있다. 그런데 현재 이용 가능한 음성 변환 기술들은 여러 가지 한계를 가지고 있다.

[0005] 예를 들어, 현재 이용 가능한 음성 변환 기술들은, 타깃 음성의 박자, 억양, 운율, 또는 감정 등 음성의 어느 하나의 특징을 변환하는 과정에서, 화자의 목소리 등 다른 특징들이 의도치 않게 함께 변하는 문제를 완전히 해결하지 못하고 있다. 다시 말해, 행복한 말투로 표현된 음성을 슬픈 말투를 가지는 음성으로 변환하는 과정에서, 화자의 목소리가 이질적으로 변경되거나, 오디오의 품질이 훼손되는 문제가 발생한다.

[0006] 또 다른 예로, 현재 이용 가능한 음성 변환 기술들은, 셋 이상의 도메인들 사이의 상호 변환 기능을 제공하지 못하고 있다. 다시 말해, 하나의 음성 변환 모델을 이용하여, 제1 감정을 가지는 음성(예컨대 행복한 말투의 음성)을 제2 감정을 가지는 음성(예컨대 슬픈 말투의 음성)으로 변환하거나, 제2 감정을 가지는 음성을 제1 감정을 가지는 음성으로 변환하는 양방향 변환만이 가능할 뿐, 하나의 음성 변환 모델을 이용하여, 셋 이상의 서로 다른 감정들 사이를 상호 변환하는 기술은 제공되지 못하고 있다.

[0007] 화자의 목소리의 동질성을 유지한 채로, 다양한 도메인들 사이에서 음성을 변환하기 위해서는, 각각의 화자마다 별도의 음성 변환 모델을 구축하거나, 및/또는 소스 도메인과 타깃 도메인으로 이루어진 하나의 쌍마다 별도의 음성 변환 모델을 구축하는 방식의 접근이 가능하다. 하지만, 각각의 음성 변환 모델 구축을 위한 학습 데이터

의 확보와 변환 모델 구축에 드는 비용 및 시간을 고려하면, 상술한 접근 방식은 현실적인 솔루션이 되지 못한다.

[0008] 따라서, 화자를 특정하지 않은 하나의 음성 변환 모델을 통해, 변환 대상 음성의 나머지 특징들은 그대로 유지한 채로, 변환 대상 음성의 일부 특징만을 변환할 수 있는 음성 변환 방법이 요구된다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 한국등록특허 제10-2057926호 (2019.12.20. 공고)

발명의 내용

해결하려는 과제

[0010] 본 발명의 몇몇 실시예들을 통해 해결하고자 하는 기술적 과제는, 신경망 기반의 오디오 변환 모델을 이용한 오디오 변환 장치 및 그 장치에서 수행되는 방법을 제공하는 것이다.

[0011] 본 발명의 몇몇 실시예들을 통해 해결하고자 하는 다른 기술적 과제는, 변환 대상 음성 고유의 일부 특징을 유지한 채, 변환하고자 하는 타깃 특징만을 선별적으로 변환하는 오디오 변환 장치 및 그 장치에서 수행되는 방법을 제공하는 것이다.

[0012] 본 발명의 몇몇 실시예들을 통해 해결하고자 하는 또 다른 기술적 과제는, 변환 대상 음성의 화자의 목소리 특징을 유지하면서 변환 대상 음성에 표현된 화자의 감정을 제1 감정으로부터 제2 감정으로 변환하는 장치 및 그 장치에서 수행되는 방법을 제공하는 것이다.

[0013] 본 발명의 몇몇 실시예들을 통해 해결하고자 하는 또 다른 기술적 과제는, 하나의 음성 변환 모델을 이용하여 셋 이상의 서로 다른 감정들 사이를 변환하는 음성 변환 장치 및 그 장치에서 수행되는 방법을 제공하는 것이다.

과제의 해결 수단

[0014] 상기 기술적 과제를 해결하기 위한, 본 발명의 일 실시예에 따른, 신경망 기반 오디오 변환 모델을 이용하여 입력 오디오 시퀀스의 특징을 변환하는 방법은, 상기 입력 오디오 시퀀스를 나타내는 입력 오디오 데이터를 상기 오디오 변환 모델에 입력하는 단계와, 상기 입력 오디오 시퀀스의 시퀀스 레벨 특징을 나타내는 제1 잠재 변수(latent variable)를 획득하는 단계와, 상기 입력 오디오 시퀀스를 구성하는 복수의 세그먼트들 각각의 특징들을 나타내는 제2 잠재 변수를 획득하는 단계와, 상기 제1 잠재 변수를 조정하는 단계와, 상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 이용하여, 출력 오디오 시퀀스를 나타내는 출력 오디오 데이터를 획득하는 단계를 포함한다.

[0015] 일 실시예에서, 상기 입력 오디오 시퀀스는 화자(speaker)의 음성(voice)을 포함하고, 상기 제1 잠재 변수는, 상기 입력 오디오 시퀀스의 제1 특징을 적어도 부분적으로 표현하는 제3 잠재 변수 및 상기 입력 오디오 시퀀스의 제2 특징을 적어도 부분적으로 표현하는 제4 잠재 변수를 포함하고, 상기 제2 잠재 변수는, 상기 입력 오디오 시퀀스에 포함된 언어적 컨텐츠를 적어도 부분적으로 표현하는 것이며, 상기 제1 잠재 변수를 조정하는 단계는, 상기 제4 잠재 변수를 그대로 유지한채 상기 제3 잠재 변수를 조정하는 단계를 포함할 수 있다.

[0016] 일 실시예에서, 상기 제1 특징은 상기 입력 오디오 시퀀스에 표현된 상기 화자의 감정을 가리키는 것이고, 상기 제2 특징은 상기 화자의 목소리 특징을 가리키는 것일 수 있다.

[0017] 일 실시예에서, 상기 제3 잠재 변수는 벡터로 표현되는 값이고,상기 제3 잠재 변수를 조정하는 단계는, 제1 감정에 대응되는 임베딩 벡터를 제2 감정에 대응되는 임베딩 벡터로 변환하는 감정 변환 함수를 상기 제3 잠재 변수에 적용하는 단계를 포함할 수 있다.

[0018] 일 실시예에서, 상기 제3 잠재 변수를 조정하는 단계는, 상기 제1 감정에 대응되는 임베딩 벡터 및 상기 제2 감정에 대응되는 임베딩 벡터를 직교화(orthogonalization)하는 단계를 더 포함할 수 있다.

[0019] 일 실시예에서, 상기 출력 오디오 시퀀스는, 상기 입력 오디오 시퀀스의 언어적 컨텐츠 및 상기 입력 오디오 시퀀스의 화자의 음성의 특징을 유지한 채, 상기 입력 오디오 시퀀스에 표현된 상기 화자의 감정이 변환된 오디오일 수 있다.

[0020] 일 실시예에서, 상기 제1 잠재 변수를 획득하는 단계는, 상기 오디오 변환 모델의 인코딩부에 의해, 상기 입력 오디오 데이터를 이용하여 상기 제3 잠재 변수 예측 분포를 획득하는 단계와, 상기 제3 잠재 변수 예측 분포로부터 상기 제3 잠재 변수를 샘플링하는 단계를 포함할 수 있다.

[0021] 일 실시예에서, 상기 제3 잠재 변수 예측 분포는 정규 분포에 해당할 수 있다.

[0022] 일 실시예에서, 상기 제1 잠재 변수를 획득하는 단계는, 상기 인코딩부에 의해, 상기 입력 오디오 데이터를 이용하여 상기 제4 잠재 변수 예측 분포를 획득하는 단계와, 상기 제4 잠재 변수 예측 분포로부터 상기 제4 잠재 변수를 샘플링하는 단계를 더 포함할 수 있다.

[0023] 일 실시예에서, 상기 제2 잠재 변수를 획득하는 단계는, 상기 입력 오디오 데이터 및 상기 제1 잠재 변수를 이용하여, 상기 제2 잠재 변수를 획득하는 단계를 포함할 수 있다.

[0024] 일 실시예에서, 상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 이용하여, 출력 오디오 데이터를 획득하는 단계는, 상기 오디오 변환 모델의 디코딩부에 의해, 상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 결합(concatenate)한 값을 이용하여, 상기 출력 오디오 데이터를 예측하기 위한 분포를 획득하는 단계와, 상기 출력 오디오 데이터를 예측하기 위한 상기 분포로부터 상기 출력 오디오 데이터를 샘플링하는 단계를 포함할 수 있다.

[0025] 일 실시예에서, 상기 출력 오디오 데이터를 보코더에 입력하여 출력 오디오 시퀀스를 합성하는 단계를 더 포함할 수 있다.

[0026] 일 실시예에서, 상기 입력 오디오 시퀀스를 단시간 푸리에 변환(Short Time Fourier Transform)하여 상기 입력 오디오 시퀀스에 대한 스펙트로그램(spectrogram) 데이터를 획득하는 단계와, 신경망 보코더를 이용하여 상기 출력 오디오 데이터로부터 상기 출력 오디오 시퀀스를 합성하는 단계를 더 포함할 수 있다.

[0027] 상술한 기술적 과제를 해결하기 위한, 본 발명의 다른 일 실시예에 따른 오디오 변환 장치는, 인코딩부, 잠재 변수 변환부, 및 디코딩부를 포함하는 오디오 변환 모델과, 출력 오디오 데이터로부터 출력 오디오 시퀀스를 합성하는 오디오 합성부를 포함한다. 상기 인코딩부는, 입력 오디오 시퀀스를 전처리한 입력 오디오 데이터로부터, 상기 입력 오디오 시퀀스의 시퀀스 레벨 특징을 나타내는 제1 잠재 변수를 인코딩하여, 상기 제1 잠재 변수를 상기 잠재 변수 변환부에 제공하고, 상기 입력 오디오 데이터 및 상기 제1 잠재 변수로부터, 상기 입력 오디오 시퀀스를 구성하는 복수의 세그먼트들 각각의 특징들을 나타내는 제2 잠재 변수를 인코딩하여, 상기 제2 잠재 변수를 상기 디코딩부에 제공하며, 상기 잠재 변수 변환부는, 상기 제1 잠재 변수를 조정하고, 조정된 제1 잠재 변수를 상기 디코딩부에 제공하고, 상기 디코딩부는, 상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 이용하여, 상기 출력 오디오 데이터를 제공한다.

[0028] 일 실시예에서, 상기 인코딩부는, 상기 입력 오디오 데이터로부터 상기 제1 잠재 변수의 평균을 인코딩하고, 상기 제1 잠재 변수의 평균에 기초하여 결정되는 예측 분포로부터 상기 제1 잠재 변수를 샘플링하며, 상기 입력 오디오 데이터 및 상기 제1 잠재 변수로부터 상기 제2 잠재 변수의 평균을 인코딩하고, 상기 제2 잠재 변수의 평균에 기초하여 결정되는 예측 분포로부터 상기 제2 잠재 변수를 샘플링 할 수 있다.

[0029] 일 실시예에서, 상기 입력 오디오 시퀀스는 화자(speaker)의 음성(voice)을 포함하고, 상기 제1 잠재 변수는, 상기 입력 오디오 시퀀스의 제1 특징을 적어도 부분적으로 표현하는 제3 잠재 변수 및 상기 입력 오디오 시퀀스의 제2 특징을 적어도 부분적으로 표현하는 제4 잠재 변수를 포함하고,상기 제2 잠재 변수는, 상기 입력 오디오 시퀀스에 포함된 언어적 컨텐츠를 적어도 부분적으로 표현하는 것이며, 상기 잠재 변수 변환부는, 상기 제4 잠재 변수를 그대로 유지한채 상기 제3 잠재 변수를 조정 할 수 있다.

[0030] 일 실시예에서, 상기 제1 특징은 상기 입력 오디오 시퀀스에 표현된 상기 화자의 감정을 가리키는 것이고, 상기 제2 특징은 상기 화자의 목소리 특징을 가리킬 수 있다.

[0031] 일 실시예에서, 상기 제3 잠재 변수는 벡터로 표현되는 값이고, 상기 잠재 변수 변환부는, 제1 감정에 대응되는 임베딩 벡터를 제2 감정에 대응되는 임베딩 벡터로 변환하는 감정 변환 함수를 상기 제3 잠재 변수에 적용하여 상기 제3 잠재 변수를 조정할 수 있다.

[0032] 일 실시예에서, 상기 디코딩부는, 상기 조정된 제1 잠재 변수 및 상기 제2 잠재 변수를 결합(concatenate)한 값을 이용하여, 상기 출력 오디오 데이터 예측 분포를 획득하고, 상기 출력 오디오 데이터 예측 분포로부터 상기 출력 오디오 데이터를 샘플링 할 수 있다.

[0033] 상술한 기술적 과제를 해결하기 위한, 본 발명의 다른 일 실시예에 따른, 컴퓨팅 장치가 신경망 기반 오디오 변환 모델을 트레이닝 하는 방법은, 복수의 오디오 시퀀스 및 상기 복수의 오디오 시퀀스 각각에 나타난 화자의 감정을 가리키는 레이블을 포함하는 학습 데이터 세트를 획득하는 단계와, 상기 학습 데이터 세트를 이용하여 상기 오디오 변환 모델의 파라미터를 업데이트하는 단계를 포함한다. 상기 오디오 변환 모델의 인코딩부는, 입력 오디오 시퀀스를 인코딩하여 상기 입력 오디오 시퀀스에 나타난 화자의 감정을 나타내는 제1 잠재 변수(latent variable) 및 상기 입력 오디오 시퀀스를 구성하는 복수의 세그먼트들 각각의 특징들을 나타내는 제2 잠재 변수를 제공하고, 상기 오디오 변환 모델의 디코딩부는, 상기 제1 잠재 변수 및 상기 제2 잠재 변수를 디코딩하여 출력 오디오 시퀀스를 제공하며, 상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는, 제1 입력 오디오 시퀀스를 상기 오디오 변환 모델에 입력할 경우 상기 제1 입력 오디오 데이터와 동일한 오디오 시퀀스가 출력될 확률이 증가하도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 포함하고, 상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는, 상기 인코딩부에 의해 제공되는 상기 제1 잠재 변수 및 상기 제2 잠재 변수의 분포가 정규 분포에 가까워지도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 포함한다.

[0034] 일 실시예에서, 상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는, 동일한 감정을 가리키는 레이블을 가지는 복수의 오디오 시퀀스를 인코딩하여 획득되는 상기 제1 잠재 변수들의 분포가 정규 분포에 가까워지도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 더 포함할 수 있다.

[0035] 일 실시예에서, 상기 오디오 변환 모델의 파라미터를 업데이트하는 단계는, 각각 서로 다른 감정을 가리키는 레이블을 가지는 복수의 오디오 시퀀스들을 인코딩하여 획득되는 상기 제1 잠재 변수들의 거리가 서로 멀어지도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 단계를 더 포함할 수 있다.

도면의 간단한 설명

[0036] 도 1은 본 발명의 일 실시예에 따른 오디오 감정 변환 장치의 입력 및 출력을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 오디오 감정 변환 장치를 나타내는 예시적인 블록도이다.

도 3은 도 2를 참조하여 설명된 오디오 감정 변환 장치의 음성 전처리부를 설명하기 위한 도면이다.

도 4는 도 2를 참조하여 설명된 오디오 감정 변환 장치의 음성 합성부를 설명하기 위한 도면이다.

도 5는 도 2를 참조하여 설명된 오디오 감정 변환 장치의 음성 변환부를 나타내는 예시적인 블록도이다.

도 6은 도 5를 참조하여 설명된 음성 변환부의 음성 변환 모델을 설명하기 위한 도면이다.

도 7은 본 발명의 몇몇 실시예에 따른 감정 변환 과정에서의 사용되는 임베딩 벡터들의 직교화(orthogonalization) 과정을 설명하기 위한 도면이다.

도 8은 도 6을 참조하여 설명된 음성 변환 모델의 인코딩부를 설명하기 위한 도면이다.

도 9는 도 6을 참조하여 설명된 음성 변환 모델의 디코딩부를 설명하기 위한 도면이다.

도 10은 본 발명의 일 실시예에 따라 음성 변환 모델을 구축하고 이를 통해 음성을 변환하는 일련의 과정을 나타내는 예시적인 흐름도이다.

도 11은 본 발명의 일 실시예에 따라 음성을 변환하는 방법을 나타내는 예시적인 흐름도이다.

도 12는 본 발명의 몇몇 실시예들에 따른 오디오 변환 장치를 구현할 수 있는 예시적인 컴퓨팅 장치를 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0037] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예들을 상세히 설명한다. 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명의 기술적 사상은 이하의 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현

될 수 있으며, 단지 이하의 실시예들은 본 발명의 기술적 사상을 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명의 기술적 사상은 청구항의 범주에 의해 정의될 뿐이다.

[0038] 각 도면의 구성요소들에 참조부호를 부가함에 있어서, 동일한 구성요소들에 대해서는 비록 다른 도면상에 표시되더라도 가능한 한 동일한 부호를 가지도록 하고 있음에 유의해야 한다. 또한, 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다.

[0039] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있다. 또 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다. 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다.

[0040] 또한, 본 발명의 구성 요소를 설명하는 데 있어서, 제1, 제2, A, B, (a), (b) 등의 용어를 사용할 수 있다. 이러한 용어는 그 구성 요소를 다른 구성 요소와 구별하기 위한 것일 뿐, 그 용어에 의해 해당 구성 요소의 본질이나 차례 또는 순서 등이 한정되지 않는다. 어떤 구성 요소가 다른 구성요소에 "연결", "결합" 또는 "접속"된다고 기재된 경우, 그 구성 요소는 그 다른 구성요소에 직접적으로 연결되거나 또는 접속될 수 있지만, 각 구성 요소 사이에 또 다른 구성 요소가 "연결", "결합" 또는 "접속"될 수도 있다고 이해되어야 할 것이다.

[0041] 명세서에서 사용되는 "포함한다 (comprises)" 및/또는 "포함하는 (comprising)"은 언급된 구성 요소, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성 요소, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0042] 본 명세서에 대한 설명에 앞서, 본 명세서에서 사용되는 몇몇 용어들에 대하여 명확하게 하기로 한다.

[0043] 본 명세서에서, 시퀀스(sequence)란 순서를 가지는 일련의 데이터들의 모음 또는 선형적으로 배열될 수 있는 일련의 데이터들의 모음을 의미한다. 본 명세서에서, 세그먼트(segment)란 시퀀스를 구성하는 단위를 가리킨다. 예를 들어, 고유한 길이를 가지는 오디오 파일 또는 오디오 데이터는 본 명세서에서 오디오 시퀀스로 이해될 수 있으며, 소정의 길이를 가지는 단위 시간으로 분절한 오디오 시퀀스의 조각은 본 명세서에서 오디오 세그먼트로 이해될 수 있다. 본 명세서에서 오디오 시퀀스를 구성하는 프레임(frame)은 오디오 시퀀스를 구성하는 세그먼트와 같은 의미로 사용된다.

[0044] 본 명세서에서, 음성 또는 발화의 언어적 특징(linguistic feature) 또는 언어적 컨텐츠(linguistic content)란, 음성에 담긴, 화자가 전달하고자 하는 메시지를 의미한다. 예를 들어 음성이 기록된 오디오로부터 식별될 수 있는 텍스트나 음소 등이 언어적 특징들에 해당한다.

[0045] 본 명세성에서, 음성의 음향적 특징(acoustic feature)이란, 화자의 음색, 높낮이, 억양 등의 특징들을 의미한다.

[0046] 본 명세서에서, 음성의 감정 특징(emotion feature)은 음향적 특징의 일부로 이해될 수 있다. 음성의 감정 특징은, 화자의 음성에 표현된 행복함, 기쁨, 슬픔, 분노, 공포 등의 감정에 따라 달라지는 음향적 특징들을 가리킨다. 언어 고유의 특성 및 화자 고유의 특성에 따라, 음성에 감정이 표현되는 방식은 다양할 수 있다. 예를 들어, 표현하고자 하는 감정에 따라, 음성을 구성하는 음의 높이, 운율, 속도 등이 달라질 수 있다. 본 명세서에서 음성에 표현된 감정 특징이란 상술한 음성 특징들을 포괄하는 의미로 해석될 수 있다.

[0047] 본 명세서에서, 음성에 표현된 화자의 목소리 특징이란, 음향적 특징의 일부로 이해될 수 있다. 음성에 표현된 화자의 목소리 특징이란, 화자 고유의 목소리를 식별시키는 특징들을 가리킨다. 화자의 목소리 특징은, 화자 고유의 발성 기관의 구조 및 발성 습관 등에 따라 달라지는 음색 및 운율적 특징들을 포괄한다.

[0048] 본 명세서에서, 음성에 표현된 감정 특징과 화자 목소리 특징은 서로 구별되지만 전적으로 상호 배타적인 것은 아니다. 예를 들어 운율과 같은 특징들은 표현하고자 하는 감정에 따라 달라지기도 하고, 화자의 말투에 따라 달라지기도 한다.

[0049] 이하, 본 발명의 몇몇 실시예들에 대하여 첨부된 도면에 따라 상세하게 설명한다.

[0050] 도 1은 본 발명의 일 실시예에 따른 오디오 감정 변환 장치(10)의 입력 및 출력을 설명하기 위한 도면이다.

[0051] 도 1에 도시된 바와 같이, 오디오 감정 변환 장치(10)는 제1 감정으로 표현된 오디오 시퀀스를 획득하여, 제2 감정으로 표현된 오디오 시퀀스를 출력하는 컴퓨팅 장치이다.

[0052] 상기 컴퓨팅 장치는 노트북, 데스크톱(desktop), 랩탑(laptop) 등이 될 수 있으나, 이에 국한되는 것은 아니며 컴퓨팅 기능이 구비된 모든 종류의 장치를 포함할 수 있다. 상기 컴퓨팅 장치의 일 예는 도 12를 더 참조하도록 한다.

[0053] 도 1은 오디오 감정 변환 장치(10)가 단일 컴퓨팅 장치로 구현된 것을 예로써 도시하고 있으나, 오디오 감정 변환 장치(10)의 제1 기능은 제1 컴퓨팅 장치에서 구현되고, 제2 기능은 제2 컴퓨팅 장치에서 구현될 수도 있다.

[0054] 본 발명의 다양한 실시예들에 따르면, 오디오 감정 변환 장치(10)는 음성 변환 모델을 구축하고, 음성 변환 모델을 이용하여, 제1 감정으로 표현된 오디오 시퀀스(1)로부터 제2 감정으로 표현된 오디오 시퀀스(3)를 생성할 수 있다.

[0055] 음성 변환 모델은, 예를 들어 인공 신경망을 기반으로 하는 모델일 수 있다. 예컨대, 심층 신경망(DNN: Deep Neural Network), 선형 신경망(Linear Neural Network), 순환 신경망(RNN: Recurrent Neural Network), 장단기 메모리(LSTM: Long Short-Term Memory), 컨볼루션 신경망(CNN: Convolutional Neural Networks) 등과 같은 모델 또는 이들을 조합하여 구성한 신경망 모델이 음성 변환 모델로서 사용될 수 있으나, 이에 한정되지는 않는다.

[0056] 음성 변환 모델은, 학습용 오디오 시퀀스 샘플들과 각 오디오 시퀀스 샘플에 대응되는 감정 레이블 데이터로 구성된, 학습용 데이터셋을 이용하여, 지도 방식(supervised learning)으로 학습될 수 있다.

[0057] 예를 들어, 행복함, 기쁨, 슬픔, 분노, 공포의 다섯 가지 감정들 사이를 변환하는 음성 변환 모델을 구축할 경우, 전문 성우 등을 통해 상기 다섯 가지 서로 다른 감정들을 각각 연기한, 충분한 수의 오디오 시퀀스 샘플들을 생성하고, 각각의 감정 레이블이 태깅된 상기 오디오 샘플들을 이용하여 음성 변환 모델이 학습될 수 있다. 또한, TV 및 라디오 프로그램 등 기존에 존재하는 오디오 시퀀스 샘플들에 감정 레이블을 사람이 태깅한 결과물을 이용하여 음성 변환 모델이 학습될 수 있다.

[0058] 음성 변환 모델의 학습 과정에서, 음성 변환 모델에 입력된 오디오 시퀀스 샘플과 가급적 동일하거나 유사한 오디오 시퀀스가 음성 변환 모델로부터 출력되도록, 음성 변환 모델의 파라미터들을 조정함으로써, 음성 변환 모델의 학습이 수행될 수 있다.

[0059] 음성 변환 모델의 학습 과정에서는, 음성 변환 모델 내부의 감정 임베딩 매트릭스가 구축된다. 학습이 진행되는 동안에, 동일한 감정 레이블을 가지는 오디오 시퀀스 샘플들에 의해 생성되는 임베딩 벡터들은 잠재 벡터 공간에서 서로 근접하게 위치하도록, 상이한 감정 레이블을 가지는 오디오 시퀀스 샘플들에 의해 생성되는 임베딩 벡터들은 잠재 벡터 공간에서 서로 멀리 위치하도록, 음성 변환 모델 내부의 감정 임베딩 매트릭스가 업데이트 된다.

[0060] 본 발명의 다양한 실시예들에 따른 음성 변환 모델의 학습 방법에 대해서는 후술하기로 한다.

[0061] 이하에서는 본 발명의 일 실시예에 따른 오디오 감정 변환 장치(10)의 기능적인 구성에 대하여 도 2를 참조하여 설명한다.

[0062] 도 2에 도시된 바와 같이, 오디오 감정 변환 장치(10)는 음성 전처리부(21), 음성 변환부(23), 음성 합성부(25), 및 저장부(27)를 포함할 수 있다. 다만, 도 2에는 본 발명의 실시예와 관련 있는 구성 요소들만이 도시되어 있다. 따라서, 본 발명이 속한 기술분야의 통상의 기술자라면 도 2에 도시된 구성요소들 외에 다른 범용적인 구성 요소들이 더 포함될 수 있음을 알 수 있다. 또한, 도 2에 도시된 오디오 감정 변환 장치(10)의 각각의 구성 요소들은 기능적으로 구분되는 기능 요소들을 나타낸 것으로서, 복수의 구성 요소가 실제 물리적 환경에서는 서로 통합되는 형태로 구현될 수도 있음에 유의한다. 이하, 각 구성요소에 대하여 설명한다.

[0063] 설명의 편의를 위하여, 이하에서는 도 2 내지 도 4를 참조하여 음성 전처리부(21), 음성 합성부(25), 및 저장부(27)를 먼저 설명한 후, 이어서 도 5 내지 도 9를 참조하여 음성 변환부(23)를 설명하기로 한다.

[0064] 음성 전처리부(21)는, 오디오 감정 변환 장치(10)에 입력된, 제1 감정으로 표현된 오디오 시퀀스(1)를 변환하기에 앞서, 상기 오디오 시퀀스(1)를 후술할 음성 변환 모델이 처리하기 적합한 형태의 데이터로 전처리한다.

[0065] 도 3을 참조하면, 몇몇 실시예에 따른 음성 전처리부(21)는 단시간 푸리에 변환(Short Time Fourier Transform)을 수행하는 STFT 변환부(31)를 포함할 수 있다. STFT 변환부(31)는 디지털 오디오 데이터(예컨대

wav 형식의 오디오)를 스펙트로그램 형태의 데이터로 변환하는 전처리 기능을 수행할 수 있다. 가령, STFT 변환부(31)는 STFT(Short Time Fourier Transform) 신호 처리를 수행하여, 오디오 감정 변환 장치(10)에 입력된, 제1 감정으로 표현된 오디오 시퀀스(1)를 스펙트로그램 데이터(35)로 변환하거나, 나아가 상기 스펙트로그램 데이터를 멜-스케일(mel-scale)로 변환할 수 있다. 멜-스케일 스펙트로그램의 경우 높은 주파수대의 정보보다 낮은 주파수대의 정보에 상대적으로 더 큰 중요도를 부여한다. 멜-스케일의 스펙트로그램은, 사람의 청각이 더 민감하게 반응하는 낮은 주파수 대의 정보에 더 큰 중요도를 부여하므로, 사람이 인식하는 음성의 특징을 보다 잘 표현해 낼 수 있다.

[0066] 음성 전처리부(21)는 STFT 변환부(31)에 의해 출력된 상기 스펙트로그램 데이터(35)를 소정의 길이를 가지는 프레임(또는 세그먼트) 단위로 분절하여, 음성 변환부(23)로 제공할 수 있다. 몇몇 실시예에서, 상기 스펙트로그램 데이터(35)는, 예컨대 하나의 프레임 당 80차원의 벡터로 표현된 데이터일 수 있다.

[0067] 음성 합성부(25)에 대해서는 도 4를 참조하여 설명한다. 음성 합성부(25)는, 음성 변환부(23)로부터 출력된 스펙트로그램 데이터(41)를 다시 디지털 오디오 데이터 형태로 변환한다. 몇몇 실시예에서, 상기 스펙트로그램 데이터(41)는 예컨대 하나의 프레임 당 80차원의 벡터로 표현된 데이터일 수 있다. 음성 합성부(25)는 스펙트로그램 데이터(41)를 이용하여 오디오 또는 음성을 합성하는 보코더일 수 있다. 상기 변환 기능을 수행할 수 있다면, 음성 합성부(25)는 어떠한 방식으로 구현되더라도 무방하다. 가령, 음성 합성부(25)는 당해 기술 분야에서 널리 알려진 보코더 모듈을 이용하여 구현될 수 있다. 예를 들어, 음성 합성부(25)는 그리핀-림 알고리즘(Griffin-Lim algorithm)을 이용한 음성 합성 모듈, 또는 WaveNet 및 WaveGlow 등 신경망 기반의 보코더 모듈을 이용하여 구현될 수 있다. 본 발명의 논지를 흐리지 않기 위해 음성 합성부(25)에 대한 더 이상의 설명은 생략하도록 한다.

[0068] 다시 도 2를 참조하여, 저장부(27)에 대하여 설명한다. 저장부(27)는 각종 데이터를 저장하고 관리한다. 특히 저장부(27)는 후술할 음성 변환 모델의 학습에 사용되는 학습용 데이터셋을 저장할 수 있다. 학습용 데이터셋은 학습용 오디오 시퀀스 샘플들과 각 오디오 시퀀스 샘플에 대응되는 감정 레이블 데이터를 포함할 수 있다. 또한, 저장부(27)는 후술할 음성 변환 모델을 구성하는 하나 이상의 신경망들에 관한 각종 파라미터 및 설정들을 저장하고 관리할 수 있다. 또한 저장부(27)는 후술할 음성 변환 모델이 감정의 변환에 사용하는 감정 임베딩 벡터들로 구성된 감정 임베딩 매트릭스를 저장하고 관리할 수 있다.

[0069] 지금까지 도 2 내지 도 4를 참조하여, 오디오 감정 변환 장치(10)의 음성 전처리부(21), 음성 합성부(25), 및 저장부(27)에 대하여 설명하였다. 이하에서는 도 5 내지 도 9를 참조하여, 오디오 감정 변환 장치(10)의 음성 변환부(25)에 대하여 상세히 설명하기로 한다.

[0070] 도 5를 참조하면, 본 발명의 일 실시예에 따른 음성 변환부(23)는, 학습부(51), 음성 변환 모델(53), 및 변환부(55)를 포함할 수 있다. 설명의 편의를 위하여, 학습부(51) 및 변환부(55)를 먼저 설명한 후, 도 6 내지 도 9를 참조하여 음성 변환 모델(53)에 대하여 상세히 설명하기로 한다.

[0071] 학습부(51)는 학습용 데이터셋을 이용하여 음성 변환 모델(53)을 학습시킨다. 즉, 학습부(51)는 학습용 데이터셋을 이용하여 음성 변환 모델(53)의 음성 변환 품질이 최적화되도록, 음성 합성 모델(53)의 파라미터들을 업데이트함으로써, 음성 변환 모델(53)을 구축할 수 있다. 상기 학습용 데이터셋은 저장부(27)로부터 제공받을 수 있을 것이나, 본 발명의 기술적 범위가 이에 한정되는 것은 아니다. 설명의 편의를 위하여, 음성 변환부(23)의 다른 구성들 및 음성 변환 모델(53)의 신경망 구조에 대해서 설명한 후에, 도 10을 참조하여 학습부(51)에 의한 학습 과정에 관하여 설명하기로 한다.

[0072] 변환부(55)는, 학습부(51)에 의해 학습된 음성 변환 모델(53)을 이용하여, 오디오 시퀀스를 나타내는 스펙트로그램의 일부 특징을 변환한다. 보다 구체적으로, 변환부(55)는, 제1 감정으로 표현된 오디오 시퀀스에 대한 스펙트로그램을 음성 변환 모델(53)에 입력하고, 그 결과로 제2 감정으로 표현된 오디오 시퀀스를 합성하기 위한 스펙트로그램을 제공받는다. 변환부(55)가 음성 변환 모델(53)에 입력하거나 음성 변환 모델(53)로부터 제공받는 스펙트로그램은, 예를 들어 소정의 길이의 프레임 단위로 분절된 스펙트로그램 데이터로 구성될 수 있다.

[0073] 음성 변환 모델(53)은, 입력된 오디오 시퀀스의 특징들 중 적어도 일부를 변환하는 모델이다. 일 실시예에서 음성 변환 모델(53)은, 스펙트로그램 데이터의 형태로 입력받은 오디오 시퀀스에 표현된 제1 감정을 표현하는 특징들을, 제2 감정을 표현하는 특징들로 변환한다. 도 6에 도시된 바와 같이, 음성 변환 모델(53)은 인코딩부(61), 디코딩부(69), 및 잠재변수 변환부(67)를 포함할 수 있다. 몇몇 실시예에서, 음성 변환 모델(53)은 감정 임베딩 매트릭스(65)를 더 포함할 수 있다.

[0074] 본 발명의 몇몇 실시예에서 음성 변환 모델(53)은 변이형 오토인코더(VAE: Variational AutoEncoder)의 구조에 기초하여 구현될 수 있다. 또한, 후술할 음성 변환 모델(53)의 구성들을 이해함에 있어서 변이형 오토인코더의 구성들이 참조될 수 있다. 그러나 본 발명이 그러한 실시예로 한정되는 것은 아니다.

[0075] 도 6을 참조하면, 인코딩부(61)는 음성 변환 모델(53)에 입력된 스펙트로그램 데이터(35)를 인공 신경망 레이어에 입력하여, 입력 오디오 시퀀스의 특징들을 나타내는 잠재 변수들을 계산한다. 전술한 바와 같이, 스펙트로그램 데이터(35)는 예컨대 하나의 프레임(35a) 당 80차원의 벡터로 표현된 데이터일 수 있다. 몇몇 실시예에서, 상기 잠재 변수들은 32차원을 가지는 벡터들일 수 있다. 다시 말해, 인코딩부(61)는 80차원의 크기를 가지는 벡터들을 입력으로 받아서 32차원을 가지는 벡터들을 출력하는 것일 수 있다.

[0076] 상기 잠재 변수들에는, 입력 오디오 시퀀스의 시퀀스 레벨 특징들을 나타내는 잠재 변수 및 상기 입력 오디오 시퀀스를 구성하는 세그먼트들 또는 프레임들 각각의 특징(이하, "세그먼트 레벨 특징")들을 나타내는 잠재 변수가 포함된다. 여기서 입력 오디오 시퀀스의 시퀀스 레벨 특징들이란, 화자의 음색 등 시퀀스 전체에 걸쳐 공통되는 특징이거나, 시퀀스 전체적인 높낮이 및 운율의 변화, 억양, 감정 등 시퀀스 전체에 대한 판단 결과 하나로 결정되는 특징들을 가리킨다. 한편, 입력 오디오 시퀀스의 세그먼트 레벨 특징들이란, 오디오 시퀀스의 각각의 세그먼트 또는 프레임에 나타난 텍스트나 음소 등 언어적 특징(linguistic feature)들을 가리킨다.

[0077] 인코딩부(61)는, 시퀀스 레벨 특징들을 나타내는 잠재 변수들(예컨대 도 6의 Z_2, Z_3) 및 세그먼트 레벨 특징들을 나타내는 잠재 변수들(예컨대 도 6의 Z_1)을 디코딩부(69)에 제공한다. 또한 인코딩부(61)는, 시퀀스 레벨 특징들을 나타내는 잠재 변수들(Z_2, Z_3) 중 적어도 일부, 예컨대 감정을 나타내는 잠재 변수(Z_3)를 잠재 변수 변환부(67)에 제공할 수 있다.

[0078] 디코딩부(69)는 인코딩부(61)로부터 제공받은 잠재 변수들 및 잠재 변수 변환부(67)로부터 제공받은 조정된 잠재 변수를 인공 신경망 레이어에 입력하여, 출력 오디오 시퀀스를 생성하기 위한 스펙트로그램 데이터(41)를 생성한다. 상기 디코딩부(69)에 입력되는 잠재 변수들은 32차원의 벡터들일 수 있으며, 상기 디코딩부(69)가 출력하는 스펙트로그램 데이터(41)는, 예컨대 하나의 프레임(41a) 당 80차원의 벡터로 표현된 데이터일 수 있다. 상기 출력 오디오 시퀀스는, 잠재 변수 변환부(67)에 의해 잠재 변수(Z_3)가 조정됨으로써 변환된 오디오의 특징들이 반영된 오디오 시퀀스일 수 있다.

[0079] 만약, 오디오 시퀀스에 표현된 감정을 나타내는 잠재 변수가 잠재 변수 변환부(67)에 의해 조정되었다면, 출력 오디오 시퀀스는 입력 오디오 시퀀스의 내용을 그대로 유지한 채로, 입력 오디오 시퀀스에 표현되었던 감정(예컨대 기쁨)과는 다른 감정(예컨대 화남)으로 표현된 오디오 시퀀스일 수 있다. 만약, 화자의 목소리 특징(예컨대 음색)을 나타내는 잠재 변수가 잠재 변수 변환부(67)에 의해 조정되었다면, 출력 오디오 시퀀스는 입력 오디오 시퀀스의 내용을 그대로 유지한 채로, 입력 오디오 시퀀스의 화자의 목소리가 다른 목소리로 변경된 오디오 시퀀스일 수 있다.

[0080] 감정 임베딩 매트릭스(63)는, 서로 다른 각각의 감정을 나타내는 임베딩 벡터들로 구성된 매트릭스다. 감정 임베딩 매트릭스(63)는, 각기 다른 하나의 감정마다 하나의 임베딩 벡터를 포함한다. 예를 들어, 오디오 감정 변환 장치(10)가, 행복함, 기쁨, 슬픔, 분노, 공포의 다섯 가지 감정들 사이를 변환하는 장치라면, 감정 임베딩 매트릭스(63)에는, 총 다섯개의 임베딩 벡터들이 포함된다.

[0081] 감정 임베딩 매트릭스(63)는, 음성 변환 모델(53)에 의한 감정 변환 과정에서, 원 감정(또는 출발 감정)을 나타내는 임베딩 벡터 및 대상 감정(또는 목적 감정)을 나타내는 임베딩 벡터를 잠재 변수 변환부(67)에 제공한다.

[0082] 감정 임베딩 벡터들로 구성되는 감정 임베딩 매트릭스(63)는, 학습용 데이터셋을 이용한 음성 변환 모델(53)의 지도 학습 과정에서 구축될 수 있다. 제1 감정에 대응되는 임베딩 벡터는, 제1 감정의 레이블이 태깅된 학습용 오디오 시퀀스들을 상기 음성 변환 모델(53)에 투입한 결과 획득되는, 감정을 나타내는 잠재 변수의 평균을 획득함으로써 결정될 수 있다. 예를 들어, 행복함이라는 감정에 대응되는 임베딩 벡터는, 행복함이 표현된 것으로 레이블링된 다수의 오디오 시퀀스 샘플들을 상기 음성 변환 모델(53)에 투입하면서 상기 음성 변환 모델(53)을 학습시킨 결과로서 얻어질 수 있다. 구체적으로, 행복함이 표현된 다수의 오디오 시퀀스 샘플들을 상기 학습된 음성 변환 모델(53)에 입력한 결과 인코딩부(61)에 의해 산출되는 각각의 잠재 변수들(Z_3)이 만드는 벡터 분포의 평균 벡터(mean vector)일 수 있다.

[0083] 잠재 변수(Z_3)가 32차원를 가지는 벡터인 몇몇 실시예에서는, 감정 임베딩 벡터 역시 32차원를 가지는

벡터이다.

[0084] 몇몇 실시예에서, 감정 임베딩 매트릭스(63)에 포함된 서로 다른 감정에 대응되는 임베딩 벡터들은, 벡터 공간 내에서의 서로 간의 거리가 최대화되도록 결정될 수 있다.

[0085] 제1 감정에 대응되는 임베딩 벡터와 제2 감정에 대응되는 임베딩 벡터가 서로 유사하게 결정된다면(즉, 벡터들 사이의 거리가 가깝다면), 음성 변환 모델(53)에 의한 제1 감정으로의 변환 결과와 제2 감정으로의 변환 결과가 서로 명확하게 구별되지 않는다. 예컨대 화난 감정을 표현하고자 하는 의도로 변환한 결과물과 기쁜 감정을 표현하고자 하는 의도로 변환한 결과물 사이에 명확한 차이가 없는 문제가 발생할 수 있다. 즉, 감정 변환의 품질이 저하된다.

[0086] 본 발명의 몇몇 실시예에서는, 음성 변환 모델(53)의 지도 학습 과정에서 감정 임베딩 매트릭스(63)가 구축될 때, 감정 임베딩 매트릭스(63)에 포함된 서로 다른 감정에 대응되는 임베딩 벡터들 사이의 거리가 최대화되도록 음성 변환 모델(53)이 갱신된다. 이는 음성 변환 모델(53)의 지도 학습 과정에서 음성 변환 모델(53)의 파라미터들을 갱신할 때, 아래 수학식 1이 최대화되도록 하는 항(term)을 포함시킴으로써 달성될 수 있다.

수학식 1

[0087]
$$\sum_{i=1}^{K}\sum_{j=i+1}^{K}\left\|\mu_3^{(i)}-\mu_3^{(j)}\right\|_1$$

[0088] (여기서 K는 서로 다른 감정의 개수, $\mu_3^{(i)}$ 및 $\mu_3^{(j)}$는 서로 다른 감정에 대응되는 감정 임베딩 벡터를 가리킴)

[0089] 본 실시예에 따르면, 감정 임베딩 매트릭스(63)에 포함된 서로 다른 감정에 대응되는 임베딩 벡터들 사이의 거리가 최대화되도록 감정 임베딩 매트릭스(63)를 구축함으로써, 제1 감정을 표현하고자 하는 의도로 변환한 결과물과 제2 감정을 표현하고자 하는 의도로 변환한 결과물이 보다 뚜렷한 차이를 가지게 된다. 즉, 우수한 감정 변환 품질을 제공할 수 있게 된다.

[0090] 본 실시예에서는, 음성 변환 모델(53)이 감정 임베딩 매트릭스(63)를 포함하는 것으로 설명하였으나, 본 발명이 그러한 실시예로 한정되는 것은 아니다. 예컨대, 오디오 시퀀스의 다른 특징들은 유지한 채, 오디오 시퀀스에 표현된 화자의 억양만을 변환하기 위한 목적으로 본 발명을 구현하는 경우, 감정 임베딩 매트릭스(63) 대신에 억양 임베딩 매트릭스가 구비될 수 있다. 이 경우, 음성 변환 모델(53)의 학습 과정에서 억양 레이블 정보가 태깅된 오디오 시퀀스 샘플들을 이용하여 억양 임베딩 매트릭스가 구축될 수 있으며, 오디오 시퀀스의 억양을 제1 억양에서 제2 억양으로 변환하는 과정 중에 억양 임베딩 매트릭스로부터 억양 임베딩 벡터들이 잠재 변수 변환부(67)로 제공될 수 있다.

[0091] 다시 도 6을 참조하면, 잠재 변수 변환부(67)는 시퀀스 레벨 특징들을 나타내는 잠재 변수들(Z_2, Z_3) 중 감정을 나타내는 잠재 변수(Z_3)를 인코딩부(61)로부터 제공받아서, 감정을 변환시키기 위하여 조정된 잠재 변수($\tilde{z}_3$)를 산출한다. 이를 위하여 잠재 변수 변환부(67)는 감정 임베딩 매트릭스(63)로부터, 대상 감정(또는 목적 감정)을 나타내는 임베딩 벡터 $\mu_3^{(target)}$ 와 원 감정(또는 출발 감정)을 나타내는 임베딩 벡터 $\mu_3^{(source)}$ 를 제공받는다. 잠재 변수 변환부(67)는 조정된 잠재 변수($\tilde{z}_3$)를 디코딩부(69)에 제공한다.

[0092] 구체적으로 잠재 변수 변환부(67)는, 수학식 2와 같은 벡터 연산을 통하여 잠재 변수(Z_3)를 조정한 잠재 변수($\tilde{z}_3$)를 산출할 수 있다.

수학식 2

$$\tilde{Z}_3 = Z_3 + w^*(\mu_3^{(target)} - \mu_3^{(source)})$$

[0093]

[0094] (여기서 $\mu_3^{(target)}$ 는 대상 감정을 나타내는 임베딩 벡터, $\mu_3^{(source)}$ 는 원 감정을 나타내는 임베딩 벡터)

[0095] 잠재 변수(Z_3)로부터 조정된 잠재 변수($\tilde{Z}_3$)를 계산하는 상기 벡터 연산은, 오디오 시퀀스에 표현된 감정을 변환시킨다는 관점에서, 감정 변환 함수로 이해될 수 있다.

[0096] 전술한 바와 같이, 본 발명의 여러 실시예들에서는, 복수의 감정들 각각을 나타내는 임베딩 벡터들의 차이를 이용한 벡터 연산을 통해, 입력된 오디오 시퀀스의 감정을 나타내는 잠재 변수(Z_3)를 조정함으로써, 셋 이상의 복수의 감정들 사이를 변환하는 음성 변환 모델을 제공할 수 있다. 즉, 원 감정과 대상 감정으로 이루어진 각각의 쌍마다 별도의 음성 변환 모델을 구축하지 않고도, 셋 이상의 서로 다른 감정들 사이를 변환하는 장치를 제공할 수 있게 된다.

[0097] 본 발명의 몇몇 실시예에서, 예시적인 수학식 2로 표현된 상기 감정 변환 함수는, 대상 감정을 나타내는 임베딩 벡터 $\mu_3^{(target)}$ 및 원 감정을 나타내는 임베딩 벡터 $\mu_3^{(source)}$ 대신에, $\mu_3^{(target)}$ 및 $\mu_3^{(source)}$를 서로 직교화(orthogonalization) 한 벡터 $\tilde{\mu}_3^{(target)}$ 및 $\tilde{\mu}_3^{(source)}$를 사용할 수 있다. 이는 유사한 속성들을 공유하는 서로 다른 감정들이 보다 명확히 구별되도록 한다.

[0098] 예를 들어, 화난 감정과 행복한 감정을 표현하는 음성들은 높은 음조와 큰 목소리 등의 공통된 속성들을 가지므로, 화난 감정에 대응되는 임베딩 벡터와 행복한 감정에 대응되는 임베딩 벡터는 유사한 성분을 가질 수 있다. 이와 같은 경우, 화난 감정과 행복한 감정 사이의 상호 감정 변환 결과 사이에 명확한 차이가 없는 문제가 발생할 수 있다. 또한, 제3의 감정을 화난 감정으로 변환한 결과물과 제3의 감정을 행복한 감정으로 변환한 결과물이 서로 명확하게 구별되지 않는 문제가 발생할 수 있다.

[0099] 상술한 문제를 해결하기 위하여, 서로 다른 감정들에 대응되는 임베딩 벡터들이 가지는 유사한 성분을 제거하는 직교화 과정이 수행될 수 있다. 구체적으로, 임베딩 벡터들의 직교화는 예컨대 Gram-Schmidt 처리를 통해 수행될 수 있다. 도 7은 벡터 v_1, v_2가 Gram-Schmidt 처리를 통해 벡터 u_1, u_2로 직교화 된 결과를 개념적으로 나타내는 도면이다. 도 7에 도시된 바와 같이, 서로 유사한 성분을 가지는 벡터 v_1, v_2를 서로 직교하는 벡터 u_1, u_2로 변환하면, 벡터들이 가지는 유사한 성분이 제거되고, 벡터들 사이의 차이가 극대화된다.

[0100] 따라서, 본 발명의 몇몇 실시예에서, 대상 감정을 나타내는 임베딩 벡터 $\mu_3^{(target)}$ 및 원 감정을 나타내는 임베딩 벡터 $\mu_3^{(source)}$를 서로 직교화 함으로써, 임베딩 벡터들 사이의 유사한 성분은 제거되고, 임베딩 벡터들 사이의 차이가 더욱 뚜렷해진다. 직교화 된 감정 임베딩 벡터들을 이용하여 전술한 수학식 2의 감정 변환 함수를 계산함으로써, 서로 유사한 속성을 공유하는 상이한 감정들 사이의 감정 변환 결과가 보다 더 뚜렷하게 구별될 수 있게 된다.

[0101] 지금까지 도 6 및 도 7을 참조하여, 음성 변환 모델(53)을 설명하였다. 이하에서는 도 8과 도 9를 참조하여 음성 변환 모델(53)의 인코딩부(61)와 디코딩부(69)에 대하여 보다 자세히 설명하기로 한다.

[0102] 도 8은 몇몇 실시예에 따른 인코딩부(61)의 신경망 구조를 나타내는 도면이다.

[0103] 전술한 바와 같이, 인코딩부(61)는 스펙트로그램 데이터(35)를 입력 받아서 시퀀스 레벨 잠재 변수들(Z_2, Z_3) 및 세그먼트 레벨 잠재 변수(Z_1)를 계산 또는 예측한다. 구체적으로 인코딩부(61)는, 스펙트로그램 데이터(35)를 인공 신경망 레이어에 입력하여 각각의 잠재 변수들을 예측하기 위한 잠재 변수 예측 분포들을 결정하고, 결정된 예측 분포에서 각각의 잠재 변수들을 샘플링함으로써, 잠재 변수들을 예측한다.

[0104] 이를 위하여 인코딩부(61)는, 각각의 잠재 변수들(Z_1, Z_2, Z_3)의 분포를 결정하기 위한 각각의 신경망 레이어들(81 내지 86)과 상기 분포로부터 각각의 잠재 변수들을 샘플링하기 위한 잠재 변수 샘플링부(87)를 포함할 수 있다.

[0105] 각각의 잠재 변수들(Z_1, Z_2, Z_3)을 계산하기 위한 신경망 레이어들(81 내지 86)은, 심층 신경망(DNN: Deep Neural Network) 레이어, 선형 신경망(Linear Neural Network) 레이어, 순환 신경망(RNN: Recurrent Neural Network) 레이어, 장단기 메모리(LSTM: Long Short-Term Memory) 레이어, 컨볼루션 신경망(CNN: Convolutional Neural Networks) 레이어 등으로 구성될 수 있다. 몇몇 실시예에서, 신경망 레이어들(81 내지 86)은, 도 8에 예시적으로 도시된 바와 같이, LSTM 레이어들(81, 83, 85) 및 이들에 연결된 선형 레이어들(82a, 82b, 84a, 84b, 86a, 86b)로 구성될 수 있는데, 본 발명이 이러한 실시예로 한정되는 것은 아니다.

[0106] 설명의 편의를 위하여, 먼저 잠재 변수 Z_2 및 Z_3를 예측하기 위한 구성들을 설명하고, 이어서 잠재 변수 Z_1을 예측하기 위한 구성들을 설명한다.

[0107] 인코딩부(61)에서 시퀀스 레벨 잠재 변수 Z_2를 예측하기 위하여, 스펙트로그램 데이터(35)가 LSTM 레이어(83)에 입력된다. 전술한 바와 같이 스펙트로그램 데이터(35)는 하나의 프레임(35a) 당 80차원의 벡터로 표현된 데이터일 수 있다.

[0108] LSTM 레이어(83)의 최종 은닉층의 상태 값이 선형 레이어(84a) 및 선형 레이어(84b)로 전달된다. LSTM 레이어(83)의 최종 은닉층의 상태 값은 256 차원의 벡터일 수 있다.

[0109] 선형 레이어(84a) 및 선형 레이어(84b)에서는 최종 은닉층의 상태 값으로부터 잠재 변수 Z_2를 예측하기 위한 분포의 평균(μ_2) 및 분산(σ_2^2)이 각각 획득된다. 상기 평균(μ_2) 및 분산(σ_2^2)은 32차원, 즉 잠재 변수 Z_2와 같은 크기의 차원을 가지는 벡터들일 수 있다. 후술하겠지만, 잠재 변수 Z_2를 예측하기 위한 분포는 가우시안 정규 분포와 유사하도록 또는 가우시안 정규 분포를 따르도록, 음성 변환 모델(53)의 레이어들이 음성 변환 모델(53)의 학습 과정 중에 갱신될 수 있다.

[0110] 잠재 변수 샘플링부(87)는, 상기 선형 레이어들(84a 및 84b)로부터 제공된 상기 평균(μ_2) 및 분산(σ_2^2)으로 정의되는 가우시안 정규 분포로부터, 잠재 변수 Z_2를 예측할 수 있다.

[0111] 인코딩부(61)에서 시퀀스 레벨 잠재 변수 Z_3를 예측하기 위하여, 스펙트로그램 데이터(35)가 LSTM 레이어(85)에 입력된다. LSTM 레이어(85)의 최종 은닉층의 상태 값이 선형 레이어(86a) 및 선형 레이어(86b)로 전달된다. 선형 레이어(86a) 및 선형 레이어(86b)에서는 최종 은닉층의 상태 값으로부터 잠재 변수 Z_3를 예측하기 위한 분포의 평균(μ_3) 및 분산(σ_3^2)이 각각 획득된다. 잠재 변수 Z_3를 예측하기 위한 분포는 가우시안 정규 분포와 유사하도록 또는 가우시안 정규 분포를 따르도록, 음성 변환 모델(53)의 레이어들이 음성 변환 모델(53)의 학습 과정 중에 갱신될 수 있다. 잠재 변수 샘플링부(87)는 상기 평균(μ_3) 및 분산(σ_3^2)으로 정의되는 가우시안 정규 분포로부터 잠재 변수 Z_3를 예측할 수 있다. 몇몇 실시예에서, LSTM 레이어(83)의 최종 은닉층의 상태 값은 256 차원의 벡터이며, 상기 평균(μ_3) 및 분산(σ_3^2)은 32차원, 즉 잠재 변수 Z_3와 같은 크기의 차원을 가지는 벡터들일 수 있다.

[0112] 본 발명의 몇몇 실시예에서, 세그먼트 레벨 잠재 변수 Z_1를 예측하기 위한 구성은 시퀀스 레벨 잠재 변수 Z_2, Z_3을 예측하기 위한 구성과는 다소 차이가 있다. 도 8에 도시된 바와 같이, 세그먼트 레벨 잠재 변수 Z_1를 예측하기 위해서, 스펙트로그램 데이터(35)와 함께, 잠재 변수 샘플링부(87)에 의해 샘플링된 시퀀스 레벨 잠재 변수 Z_2, Z_3가 LSTM 레이어(81)에 입력된다.

[0113] 몇몇 실시예에서, 스펙트로그램 데이터(35)는 한 프레임(35a) 당 80차원의 벡터로 표현되는 데이터이고, 시퀀스 레벨 잠재 변수 Z_2, Z_3는 32차원의 벡터로 표현되는 데이터이다. LSTM 레이어(81)는 상기 스펙트로그램 데이터(35) 및 잠재 변수들 Z_2, Z_3를 서로 연결한(concatenation), 하나의 프레임 당 총 144차원의 크기를 가지는 벡

터로 표현되는 데이터를 입력 받는다.

[0114] LSTM 레이어(81)의 최종 은닉층의 상태 값이, 선형 레이어(82a) 및 선형 레이어(82b)로 전달된다. 선형 레이어(82a) 및 선형 레이어(82b)에서는 최종 은닉층의 상태 값으로부터 세그먼트 레벨 잠재 변수 Z_1를 예측하기 위한 분포의 평균(μ_1) 및 분산(σ_1^2)이 각각 획득된다. 잠재 변수 샘플링부(87)는 평균(μ_1) 및 분산(σ_1^2)으로 정의되는 가우시안 정규 분포로부터 잠재 변수 Z_1를 예측할 수 있다. LSTM 레이어(81)의 최종 은닉층의 상태 값은 256 차원의 벡터일 수 있고, 상기 평균(μ_1) 및 분산(σ_1^2)은 32차원, 즉 잠재 변수 Z_1와 같은 크기의 차원을 가지는 벡터들일 수 있다

[0115] 몇몇 실시예에서, 잠재 변수 샘플링부(87)는, LSTM 레이어(81) 및 선형 레이어들(82a, 82b)로부터 획득된 μ_1 및 σ_1^2을 이용하여, 정규 분포 $N(\mu_1, \sigma_1^2)$로부터 잠재 변수 Z_1을 샘플링하고, LSTM 레이어(83) 및 선형 레이어들(84a, 84b)로부터 획득된 μ_2 및 σ_2^2을 이용하여, 정규 분포 $N(\mu_2, \sigma_2^2)$로부터 잠재 변수 Z_2를 샘플링하며, LSTM 레이어(85) 및 선형 레이어들(86a, 86b)로부터 획득된 μ_2 및 σ_2^2을 이용하여, 정규 분포 $N(\mu_2, \sigma_2^2)$로부터 잠재 변수 Z_3를 샘플링한다. 이를 수식으로 나타내면 아래 수학식 3과 같다.

수학식 3

$$Z_1 \sim N(\mu_1, \sigma_1^2)$$

$$Z_2 \sim N(\mu_2, \sigma_2^2)$$

$$Z_3 \sim N(\mu_3, \sigma_3^2)$$

[0116]

[0117] 몇몇 실시예에서는, 음성 변환 모델(53)을 구성하는 신경망들의 엔드투엔드 학습을 원활하게 하기 위하여, 잠재 변수 샘플링부(87)가 잠재 변수들을 샘플링하는 방법에 재파라미터화(reparametrization) 기법을 적용하여 아래 수학식 4와 같이 구현할 수도 있다.

수학식 4

$$Z_1 = \mu_1 + \sigma_1 \epsilon, \quad where\ \epsilon \sim N(0,1)$$

$$Z_2 = \mu_2 + \sigma_2 \epsilon, \quad where\ \epsilon \sim N(0,1)$$

$$Z_3 = \mu_3 + \sigma_3 \epsilon, \quad where\ \epsilon \sim N(0,1)$$

[0118]

[0119] 지금까지 도 8을 참조하여 인코딩부(61)의 신경망 레이어 구조 및 인코딩부(61)가 잠재 변수들(Z_1, Z_2, Z_3)을 예측하는 동작에 대하여 설명하였다. 이하에서는, 도 9를 참조하여 디코딩부(69)에 대하여 설명한다.

[0120] 전술한 바와 같이, 디코딩부(69)는 잠재 변수들을 인공 신경망 레이어에 입력하여, 출력 오디오 시퀀스를 생성하기 위한 스펙트로그램 데이터(41)를 제공한다.

[0121] 이를 위하여 디코딩부(69)는, 출력 스펙트로그램 데이터(41)의 분포를 결정하기 위한 신경망 레이어들(93, 95a, 95b) 및 스펙트로그램 샘플링부(97)를 포함할 수 있다.

[0122] 전술한 신경망 레이어들(81 내지 86)과 유사하게, 신경망 레이어들(93, 95a, 95b)은, 예컨대 심층 신경망, 선형

신경망, 순환 신경망, LSTM, 컨볼루션 신경망 모델 등의 레이어들로 구성될 수 있다. 몇몇 실시예에서, 신경망 레이어들(93, 95a, 95b)은, LSTM 레이어(93) 및 그에 연결된 선형 레이어들(95a, 95b)로 구성될 수 있는데, 본 발명이 이러한 실시예로 한정되는 것은 아니다.

[0123] 디코딩부(69)는 인코딩부(61)로부터 세그먼트 레벨 잠재 변수(Z_1) 및 시퀀스 레벨 잠재 변수(Z_2)를 제공받고, 잠재 변수 변환부(67)로부터 조정된 시퀀스 레벨 잠재 변수(Z_3)를 제공받는다. 몇몇 실시예에서, 잠재 변수들(Z_1, Z_2, Z_3)은 각각 32차원의 크기를 가지는 벡터들이다.

[0124] LSTM 레이어(93)는, 상기 잠재 변수들(Z_1, Z_2, Z_3)을 서로 연결한(concatenation), 하나의 프레임 당 총 96차원의 크기를 가지는 벡터로 표현되는 데이터(91)를 입력으로 받을 수 있다.

[0125] LSTM 레이어(93)의 출력층의 시퀀스는 선형 레이어(95a, 95b)로 전달된다. 상기 출력 시퀀스는 하나의 프레임 당 256 차원을 가지는 데이터일 수 있다.

[0126] 선형 레이어(95a) 및 선형 레이어(95b)에서는, 출력 시퀀스로부터 출력 스펙트로그램 데이터(41)를 예측하기 위한 분포의 평균(μ_x) 및 분산(σ_x^2)이 각각 획득된다. 스펙트로그램 데이터(41)가 하나의 프레임(41a) 당 80차원의 벡터로 표현된 데이터인 몇몇 실시예에서, 상기 평균(μ_x) 및 분산(σ_x^2)은 80차원의 벡터로 표현된 데이터이다.

[0127] 스펙트로그램 샘플링부(97)는 상기 평균(μ_x) 및 분산(σ_x^2)에 의해 정의되는 예측 분포로부터 출력 스펙트로그램 데이터(41)를 예측할 수 있다.

[0128] 스펙트로그램 샘플링부(97)에 의해 샘플링된 출력 스펙트로그램 데이터(41)는 전술한 바와 같이 음성 합성부(25)에 제공되어 디지털 오디오 데이터 형태로 변환될 수 있다.

[0129] 지금까지 도 8 및 도 9를 참조하여 음성 변환 모델(53)의 인코딩부(61) 및 디코딩부(69)에 대하여 설명하였다. 이하에서는, 도 10 및 도 11을 참조하여, 본 발명의 다른 실시예에 따라, 음성 변환 모델을 구축하고 이를 통해 음성을 변환하는 방법을 설명한다.

[0130] 도 10은 본 발명의 일 실시예에 따라, 음성 변환 모델을 구축하고 이를 통해 음성을 변환하는 일련의 과정을 나타내는 예시적인 흐름도이다. 단, 이는 본 발명의 목적을 달성하기 위한 바람직한 실시예일 뿐이며, 필요에 따라 일부 단계가 추가되거나 삭제될 수 있음은 물론이다.

[0131] 도 10에 도시된 음성 변환 방법의 각 단계는 예컨대 오디오 감정 변환 장치(10)와 같은 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말하면, 상기 음성 변환 방법의 각 단계는 컴퓨팅 장치의 프로세서에 의해 실행되는 하나 이상의 인스트럭션들로 구현될 수 있다. 상기 음성 변환 방법에 포함되는 모든 단계는 하나의 물리적인 컴퓨팅 장치에 의하여 실행될 수도 있을 것이나, 상기 방법의 제1 단계들은 제1 컴퓨팅 장치에 의하여 수행되고, 상기 방법의 제2 단계들은 제2 컴퓨팅 장치에 의하여 수행될 수도 있다. 예컨대, 도 10에 학습 과정(단계 S100 및 S200)과 변환 과정(단계 S300 및 S400)은 서로 다른 컴퓨팅 장치에 의해 수행될 수도 있다. 이하에서는, 상기 음성 변환 방법의 각 단계가 오디오 감정 변환 장치(10)에 의해 수행되는 것을 가정하여 설명을 이어가도록 한다. 다만, 설명의 편의를 위해, 상기 음성 변환 방법에 포함되는 각 단계의 동작 주체는 그 기재가 생략될 수도 있다.

[0132] 도 10에 도시된 바와 같이, 본 실시예에 따른 음성 변환 방법은, 음성 변환 모델이 학습할 학습용 데이터셋을 획득하는 단계 S100, 학습용 데이터셋을 이용하여 음성 변환 모델을 구축하는 단계 S200, 변환하고자 하는 음성이 포함된 오디오 시퀀스를 획득하는 단계 S300, 음성 변환 모델을 이용하여 음성을 변환하는 단계 S400을 포함할 수 있다.

[0133] 먼저 단계 S100에서는, 음성 변환 모델을 학습시키기 위한 학습용 데이터셋이 획득된다.

[0134] 학습용 데이터셋은, 학습용 오디오 시퀀스 샘플들과 각 오디오 시퀀스 샘플에 대응되는 감정 레이블 데이터로 구성될 수 있다. 몇몇 실시예에서, 예컨대 행복함, 기쁨, 슬픔, 분노, 공포의 다섯 가지 감정들 사이를 변환하는 음성 변환 모델을 구축하고자 할 경우, 전문 성우 등을 통해 상기 다섯 가지 서로 다른 감정들을 각각 연기한, 충분한 수의 오디오 시퀀스 샘플들과, 각 샘플에 대응되는 감정 레이블 데이터가 획득될 수 있다. 또한, TV 및 라디오 프로그램 등 기존에 존재하는 오디오 시퀀스 샘플들에 사람이 부여한 감정 레이블 데이터가 학습용

데이터셋으로 활용될 수 있다.

[0135] 단계 S200에서는, 상기 학습용 데이터셋을 이용하여 음성 변환 모델을 구축할 수 있다. 단계 S200은, 예컨대 도 5를 참조하여 설명한 학습부(51)에 의해 수행될 수 있다. 몇몇 실시예에서, 음성 변환 모델은 도 6 내지 9를 참조하여 설명한 음성 변환 모델(53)일 수 있다. 단계 S200에서는, 상기 학습용 데이터셋에 포함된 오디오 시퀀스 샘플들과 각 샘플에 대응되는 감정 레이블 데이터를 이용하여, 상기 음성 변환 모델(53)을 지도 학습(supervised learning) 방식으로 학습시킬 수 있다.

[0136] 음성 변환 모델(53)의 학습 과정에서는, 음성 변환 모델에 입력된 오디오 시퀀스 샘플과 가급적 동일하거나 유사한 오디오 시퀀스가 음성 변환 모델(53)로부터 출력되도록, 음성 변환 모델(53)을 구성하는 신경망 레이어들의 파라미터들을 업데이트함으로써, 음성 변환 모델(53)이 학습될 수 있다. 다시 말해, 제1 입력 오디오 시퀀스를 음성 변환 모델(53)에 입력할 경우 상기 제1 입력 오디오 데이터와 동일한 오디오 시퀀스가 출력될 확률이 증가하도록 음성 변환 모델(53)의 파라미터들을 업데이트 함으로써, 음성 변환 모델(53)이 학습될 수 있다.

[0137] 또한, 음성 변환 모델(53)의 학습은, 음성 변환 모델(53)의 인코딩부(61)에서 예측되는 잠재 변수들(Z_1, Z_2, Z_3)의 예측 분포가 가우시안 정규 분포에 가까워지도록, 음성 변환 모델(53)의 파라미터들을 업데이트하는 것을 포함할 수 있다. 몇몇 실시예에서, 잠재 변수들(Z_1, Z_2, Z_3)의 예측 분포가 가우시안 정규 분포에 가까워지도록 하는 것은, 음성 변환 모델(53)의 파라미터들을 갱신할 때, 잠재 변수들(Z_1, Z_2, Z_3)의 예측 분포와 정규 분포 사이의 쿨백-라이블러 발산(KL-Divergence)이 최소화되도록 하는 항(term)을 포함시킴으로써 달성될 수 있다.

[0138] 또한, 몇몇 실시예에서, 음성 변환 모델(53)의 학습은, 서로 다른 감정을 가리키는 레이블을 가지는 복수의 오디오 시퀀스들을 인코딩부(61)가 예측하여 획득되는 잠재 변수들(Z_3)의 사이의 거리가 서로 멀어지도록 상기 오디오 변환 모델의 파라미터를 업데이트 하는 것을 포함할 수 있다. 다시 말해, 학습 과정에서 감정 임베딩 매트릭스(63)가 구축될 때, 서로 다른 감정에 대응되는 임베딩 벡터들이 벡터 공간 내에서의 서로 멀리 떨어지도록 음성 변환 모델(53)의 파라미터들이 조정된다. 이에 관해서는 감정 임베딩 매트릭스(63)의 구축에 관하여 전술한 내용이 참조될 수 있다.

[0139] 전술한 실시예에 따른 음성 변환 모델(53)의 학습 과정을 이해함에 있어서는, 변이형 오토인코더의 학습 과정이 일반적으로 참조될 수 있으나, 본 발명이 그러한 실시예로 한정되는 것은 아니다.

[0140] 다시 도 9를 참조하여, 음성 변환 모델(53)의 학습이 완료된 후, 음성 변환 모델(53)을 이용하여 음성을 변환하는 과정을 설명한다.

[0141] 단계 S300에서는 변환하고자 하는 오디오 시퀀스가 획득된다. 오디오 시퀀스는 제1 감정이 표현된 음성일 수 있으며, 단계 S400의 과정을 통해 제2 감정이 표현된 음성으로 변환될 수 있다.

[0142] 단계 S400에서는 예컨대 오디오 감정 변환 장치(10)의 음성 변환 모델(53)을 이용하여 단계 S300에서 획득된 오디오 시퀀스가 변환된다. 이하에서는 도 11을 참조하여 단계 S400의 세부 사항을 보다 자세히 설명한다.

[0143] 도 11은 본 발명의 일 실시예에 따라 예컨대 오디오 감정 변환 장치(10)가 음성을 변환하는 방법을 나타내는 예시적인 흐름도이다.

[0144] 먼저 단계 S410에서는, 입력된 오디오 시퀀스가 전처리된다. 오디오 시퀀스의 전처리는, 예컨대 오디오 감정 변환 장치(10)의 음성 전처리부(21)에 의해 수행될 수 있다. 입력된 오디오 시퀀스는 화자의 음성을 포함하는 것일 수 있다. 몇몇 실시예에서, 오디오 시퀀스의 전처리 과정은 입력된 오디오 시퀀스를 단시간 푸리에 변환(Short Time Fourier Transform) 처리하여, 멜-스케일 스펙트로그램 등의 데이터로 변환하는 과정을 포함할 수 있다. 몇몇 실시예에서, 오디오 시퀀스가 전처리 과정은 상기 멜-스케일 스펙트로그램 데이터를 소정의 길이를 가지는 프레임(또는 세그먼트) 단위로 분절하는 과정을 포함할 수 있다.

[0145] 단계 S420에서는, 상기 전처리된 스펙트로그램 데이터를 인코딩하여 상기 오디오 시퀀스의 시퀀스 레벨 특징을 나타내는 제1 잠재 변수가 획득된다. 상기 스펙트로그램 데이터의 인코딩 과정은 예컨대 오디오 감정 변환 장치(10)의 음성 변환 모델(53)을 구성하는 인코딩부(61)에 의해 수행될 수 있다. 몇몇 실시예에서, 오디오 시퀀스의 시퀀스 레벨 특징을 나타내는 제1 잠재 변수는, 입력된 오디오 시퀀스에 표현된 화자의 감정을 가리키는 특징들을 나타내는 벡터인 제3 잠재 변수 및 화자의 목소리 특징들을 가리키는 벡터인 제4 잠재 변수를 포함할 수 있다.

[0146] 단계 S420에서 제1 잠재 변수를 획득하는 과정은, 시퀀스 레벨 잠재 변수들(Z_2, Z_3)을 획득하기 위한 인코딩부(61)의 신경망 레이어들에 관하여 앞서 설명한 내용이 더 참조될 수 있다.

[0147] 단계 S430에서는, 상기 전처리된 스펙트로그램 데이터를 인코딩하여 상기 오디오 시퀀스의 세그먼트 레벨 특징을 나타내는 제2 잠재 변수가 획득된다. 상기 제2 잠재 변수의 인코딩 과정은 제1 잠재 변수의 인코딩 과정과 마찬가지로 인코딩부(61)에 의해 수행될 수 있다. 몇몇 실시예에서, 제2 잠재 변수의 인코딩 과정에는, 상기 전처리된 스펙트로그램 데이터와 함께, 상기 제1 잠재 변수가 이용될 수 있다. 몇몇 실시예에서, 오디오 시퀀스의 세그먼트 레벨 특징들을 나타내는 제2 잠재 변수는, 오디오 시퀀스의 각각의 세그먼트 또는 프레임에 나타난 텍스트나 음소 등 언어적 특징(linguistic feature) 또는 언어적 컨텐츠(linguistic contents)를 나타내는 벡터일 수 있다.

[0148] 단계 S430에서 제2 잠재 변수를 획득하는 과정은, 세그먼트 레벨 잠재 변수(Z_1)를 획득하기 위한 인코딩부(61)의 신경망 레이어들에 관하여 앞서 설명한 내용이 더 참조될 수 있다.

[0149] 단계 S440에서는, 상기 제1 잠재 변수가 조정된다. 몇몇 실시예에서, 단계 S440는, 제1 잠재 변수에 포함된 제3 잠재 변수 및 제4 잠재 변수 중에, 제3 잠재 변수만을 조정하는 것일 수 있다. 다시 말해, 몇몇 실시예에서 단계 S440는, 입력된 오디오 시퀀스의 화자의 음색 등 목소리 특징들은 그대로 유지한 채, 입력된 오디오 시퀀스에 표현된 화자의 감정만을 조정하는 것일 수 있다.

[0150] 몇몇 실시예에서, 화자의 감정을 가리키는 특징들을 나타내는 제3 잠재 변수를 조정하는 과정은, 대상 감정(또는 목적 감정)을 나타내는 제2 임베딩 벡터와 원 감정(또는 출발 감정)을 나타내는 제1 임베딩 벡터의 차이만큼 벡터를 이동시키는 감정 변환 함수를 상기 제3 잠재 변수에 적용함으로써 수행될 수 있다. 몇몇 실시예에서는, 잠재 변수 변환부(67)와 관련하여 앞서 설명한 바와 같이, 상기 제1 임베딩 벡터와 제2 임베딩 벡터 대신에, 상기 제1 임베딩 벡터와 제2 임베딩 벡터를 서로 직교화(orthogonalization) 한 벡터들이 감정 변환 함수에 이용될 수 있다.

[0151] 단계 S440에 대해서는, 잠재 변수 변환부(67)와 관련하여 앞서 설명된 내용이 더 참조될 수 있을 것이다.

[0152] 단계 S450에서는, 상기 단계 S440에서 조정된 제1 잠재 변수와 상기 단계 S430에서 획득된 제2 잠재 변수를 이용하여 출력 오디오 시퀀스를 나타내는 스펙트로그램 데이터가 획득된다. 단계 S450은, 예컨대 오디오 감정 변환 장치(10)의 음성 변환 모델(53)을 구성하는 디코딩부(69)에 의해 수행될 수 있다. 몇몇 실시예에서, 상기 조정된 제1 잠재 변수 및 제2 잠재 변수를 결합(concatenation)한 값을 이용하여 상기 출력 스펙트로그램 데이터를 예측하기 위한 분포가 획득되고, 이로부터 출력 스펙트로그램 데이터가 샘플링될 수 있다.

[0153] 단계 S450에 대해서는, 디코딩부(69)와 관련하여 앞서 설명된 내용이 더 참조될 수 있을 것이다.

[0154] 단계 S460에서는, 출력 오디오 시퀀스를 나타내는 스펙트로그램 데이터로부터 출력 오디오 시퀀스가 합성된다. 단계 S460은, 예컨대 오디오 감정 변환 장치(10)의 음성 합성부(25)에 의해 수행될 수 있으며, 음성 합성부(25)에 관하여 전술한 내용이 참조될 수 있다. 단계 S460에서는, 상기 스펙트로그램 데이터가 예컨대 신경망 기반의 보코더 모듈 등에 입력되어, 디지털 오디오 데이터 형태의 오디오 시퀀스로 합성될 수 있다.

[0155] 지금까지 도 1 내지 도 11을 참조하여, 본 발명의 몇몇 실시예들에 따른 오디오 감정 변환 방법 및 장치와, 그 응용분야에 대해서 설명하였다. 이하에서는, 본 발명의 몇몇 실시예들에 따른 오디오 감정 변환 장치(10)를 구현할 수 있는 예시적인 컴퓨팅 장치(1500)에 대하여 설명하도록 한다.

[0156] 도 12는 본 발명의 몇몇 실시예들에 따른 오디오 감정 변환 장치(10)를 구현할 수 있는 예시적인 컴퓨팅 장치(1500)를 나타내는 하드웨어 구성도이다.

[0157] 도 12에 도시된 바와 같이, 컴퓨팅 장치(1500)는 하나 이상의 프로세서(1510), 버스(1550), 통신 인터페이스(1570), 프로세서(1510)에 의하여 수행되는 컴퓨터 프로그램(1591)을 로드(load)하는 메모리(1530)와, 컴퓨터 프로그램(1591)을 저장하는 스토리지(1590)를 포함할 수 있다. 다만, 도 12에는 본 발명의 실시예와 관련 있는 구성 요소들만이 도시되어 있다. 따라서, 본 발명이 속한 기술분야의 통상의 기술자라면 도 12에 도시된 구성요소들 외에 다른 범용적인 구성 요소들이 더 포함될 수 있음을 알 수 있다.

[0158] 프로세서(1510)는 컴퓨팅 장치(1500)의 각 구성의 전반적인 동작을 제어한다. 프로세서(1510)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 또는 본 발명의 기술 분야에 잘 알려진 임의의 형태의 프로세서를 포함하여 구성될 수 있다. 또한, 프로세서

(1510)는 본 발명의 실시예들에 따른 방법을 실행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다. 컴퓨팅 장치(1500)는 하나 이상의 프로세서를 구비할 수 있다.

[0159] 메모리(1530)는 각종 데이터, 명령 및/또는 정보를 저장한다. 메모리(1530)는 본 발명의 실시예들에 따른 방법을 실행하기 위하여 스토리지(1590)로부터 하나 이상의 프로그램(1591)을 로드할 수 있다. 메모리(1530)는 RAM과 같은 휘발성 메모리로 구현될 수 있을 것이나, 본 발명의 기술적 범위가 이에 한정되는 것은 아니다.

[0160] 버스(1550)는 컴퓨팅 장치(1500)의 구성 요소 간 통신 기능을 제공한다. 버스(1550)는 주소 버스(Address Bus), 데이터 버스(Data Bus) 및 제어 버스(Control Bus) 등 다양한 형태의 버스로 구현될 수 있다.

[0161] 통신 인터페이스(1570)는 컴퓨팅 장치(1500)의 유무선 인터넷 통신을 지원한다. 또한, 통신 인터페이스(1570)는 인터넷 통신 외의 다양한 통신 방식을 지원할 수도 있다. 이를 위해, 통신 인터페이스(1570)는 본 발명의 기술 분야에 잘 알려진 통신 모듈을 포함하여 구성될 수 있다.

[0162] 몇몇 실시예들에 따르면, 통신 인터페이스(1570)는 생략될 수도 있다.

[0163] 스토리지(1590)는 상기 하나 이상의 프로그램(1591)과 각종 데이터를 비임시적으로 저장할 수 있다. 가령, 컴퓨팅 장치(1500)를 통해 텍스트 생성 장치(10)가 구현되는 경우라면, 상기 각종 데이터는 저장부(400)에 의해 관리되는 데이터를 포함할 수 있다.

[0164] 스토리지(1590)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크, 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체를 포함하여 구성될 수 있다.

[0165] 컴퓨터 프로그램(1591)은 메모리(1530)에 로드될 때 프로세서(1510)로 하여금 본 발명의 다양한 실시예에 따른 방법/동작을 수행하도록 하는 하나 이상의 인스트럭션들을 포함할 수 있다. 즉, 프로세서(1510)는 상기 하나 이상의 인스트럭션들을 실행함으로써, 본 발명의 다양한 실시예에 따른 방법/동작들을 수행할 수 있다.

[0166] 위와 같은 경우, 컴퓨팅 장치(1500)를 통해 본 발명의 몇몇 실시예들에 따른 오디오 감정 변환 장치(10)가 구현될 수 있다.

[0167] 지금까지 도 1 내지 도 12을 참조하여 본 발명의 다양한 실시예들 및 그 실시예들에 따른 효과들을 언급하였다. 본 발명의 기술적 사상에 따른 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

[0168] 지금까지 도 1 내지 도 12을 참조하여 설명된 본 발명의 기술적 사상은 컴퓨터가 읽을 수 있는 매체 상에 컴퓨터가 읽을 수 있는 코드로 구현될 수 있다. 상기 컴퓨터로 읽을 수 있는 기록 매체는, 예를 들어 이동형 기록 매체(CD, DVD, 블루레이 디스크, USB 저장 장치, 이동식 하드 디스크)이거나, 고정식 기록 매체(ROM, RAM, 컴퓨터 구비 형 하드 디스크)일 수 있다. 상기 컴퓨터로 읽을 수 있는 기록 매체에 기록된 상기 컴퓨터 프로그램은 인터넷 등의 네트워크를 통하여 다른 컴퓨팅 장치에 전송되어 상기 다른 컴퓨팅 장치에 설치될 수 있고, 이로써 상기 다른 컴퓨팅 장치에서 사용될 수 있다.

[0169] 이상에서, 본 발명의 실시예를 구성하는 모든 구성 요소들이 하나로 결합되거나 결합되어 동작하는 것으로 설명되었다고 해서, 본 발명의 기술적 사상이 반드시 이러한 실시예에 한정되는 것은 아니다. 즉, 본 발명의 목적 범위 안에서라면, 그 모든 구성요소들이 하나 이상으로 선택적으로 결합하여 동작할 수도 있다.

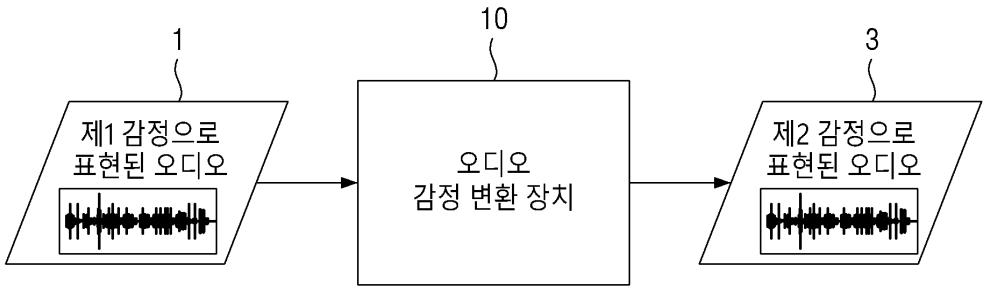
[0170] 도면에서 동작들이 특정한 순서로 도시되어 있지만, 반드시 동작들이 도시된 특정한 순서로 또는 순차적 순서로 실행되어야만 하거나 또는 모든 도시 된 동작들이 실행되어야만 원하는 결과를 얻을 수 있는 것으로 이해되어서는 안 된다. 특정 상황에서는, 멀티태스킹 및 병렬 처리가 유리할 수도 있다. 더욱이, 위에 설명한 실시예들에서 다양한 구성들의 분리는 그러한 분리가 반드시 필요한 것으로 이해되어서는 안 되고, 설명된 프로그램 컴포넌트들 및 시스템들은 일반적으로 단일 소프트웨어 제품으로 함께 통합되거나 다수의 소프트웨어 제품으로 패키지 될 수 있음을 이해하여야 한다.

[0171] 이상 첨부된 도면을 참조하여 본 발명의 실시예들을 설명하였지만, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 본 발명이 다른 구체적인 형태로도 실시될 수 있다는 것을 이해할 수 있다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적인 것이 아닌 것으로 이해해야만 한다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명에 의해 정의되는 기술적 사상의 권리범위에 포함되는 것으로

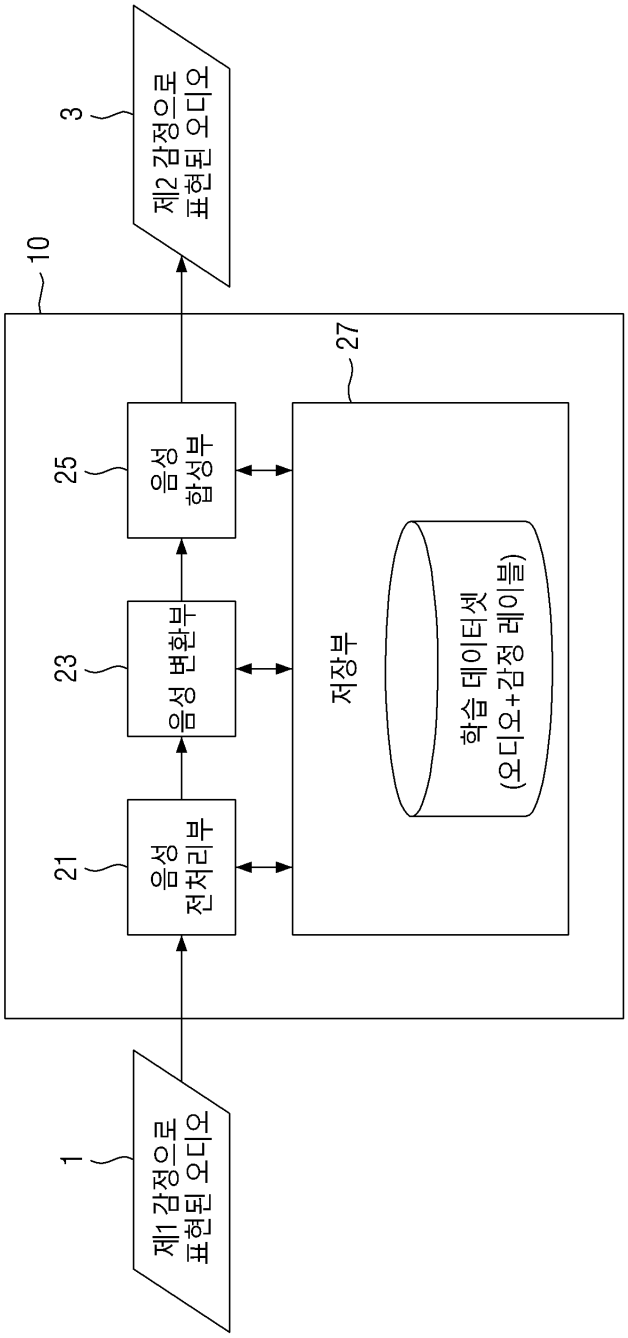
해석되어야 할 것이다.

도면

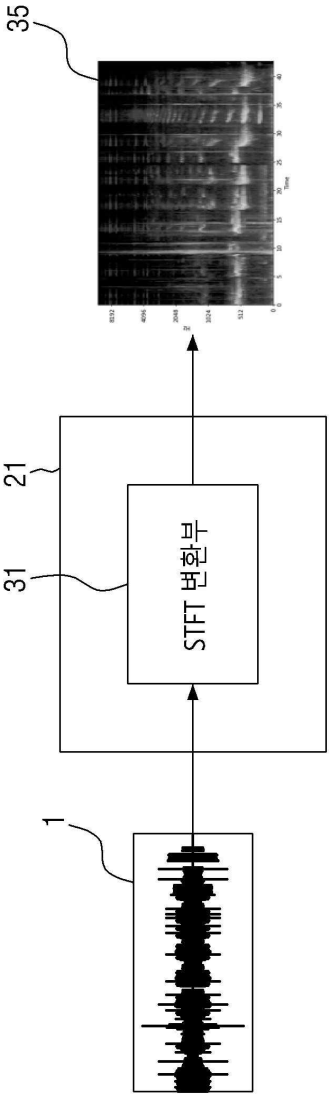
도면1



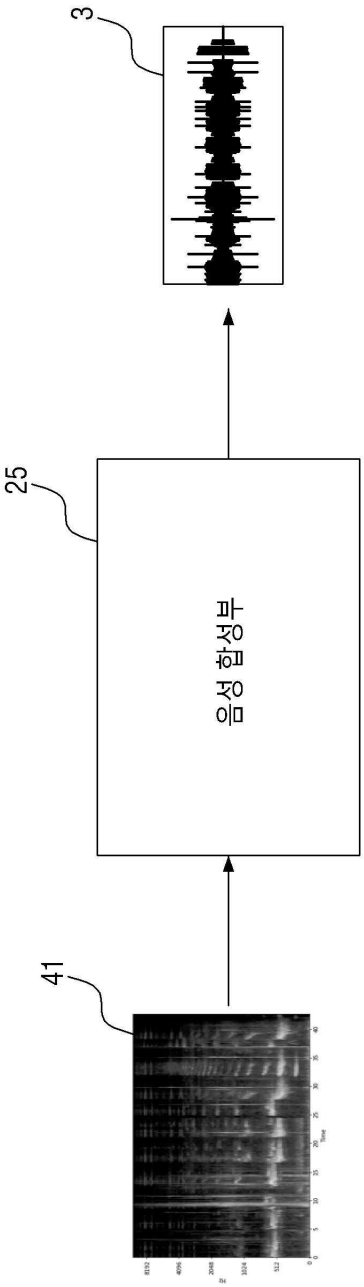
도면2


제1 감정으로
표현된 오디오
1
10
음성
전처리부
21
음성 변환부
23
음성
합성부
25
저장부
27
학습 데이터셋
(오디오+감정 레이블)
제2 감정으로
표현된 오디오
3

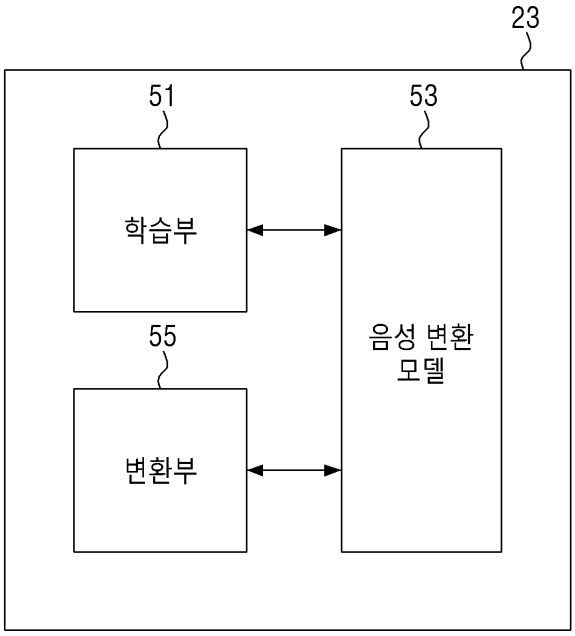
도면3



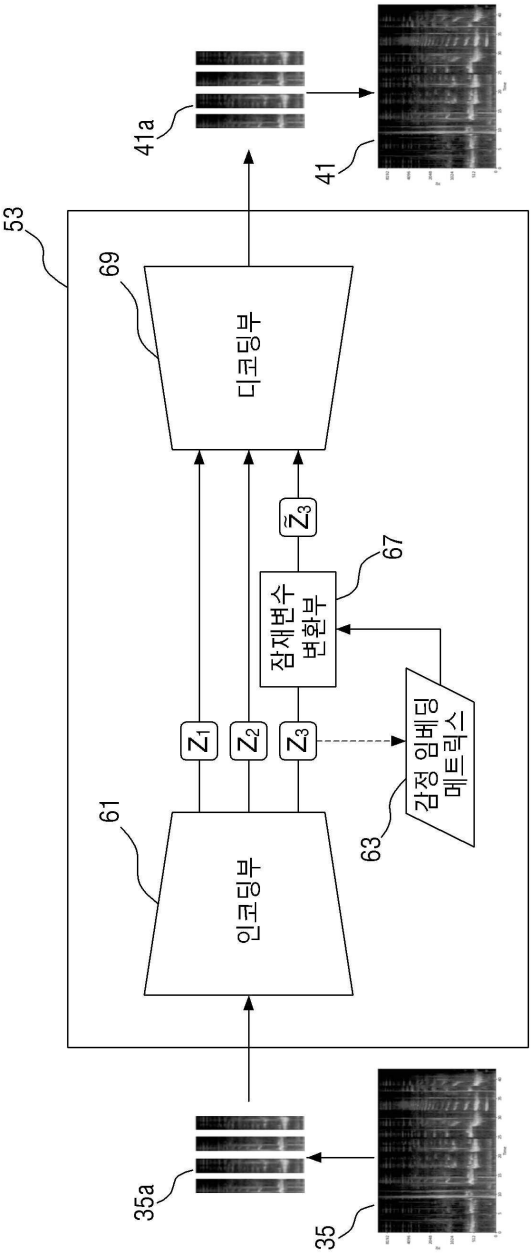
도면4



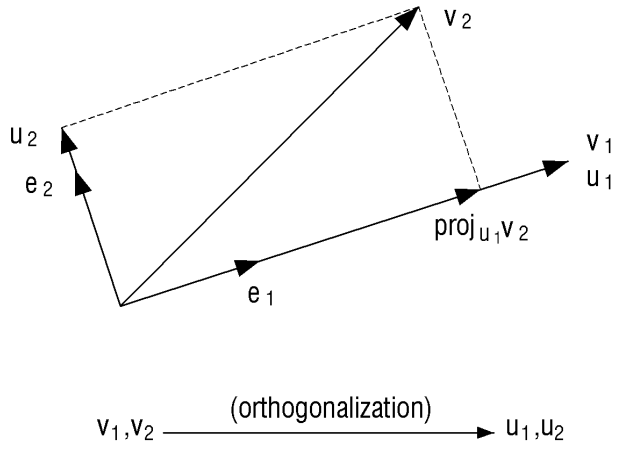
도면5



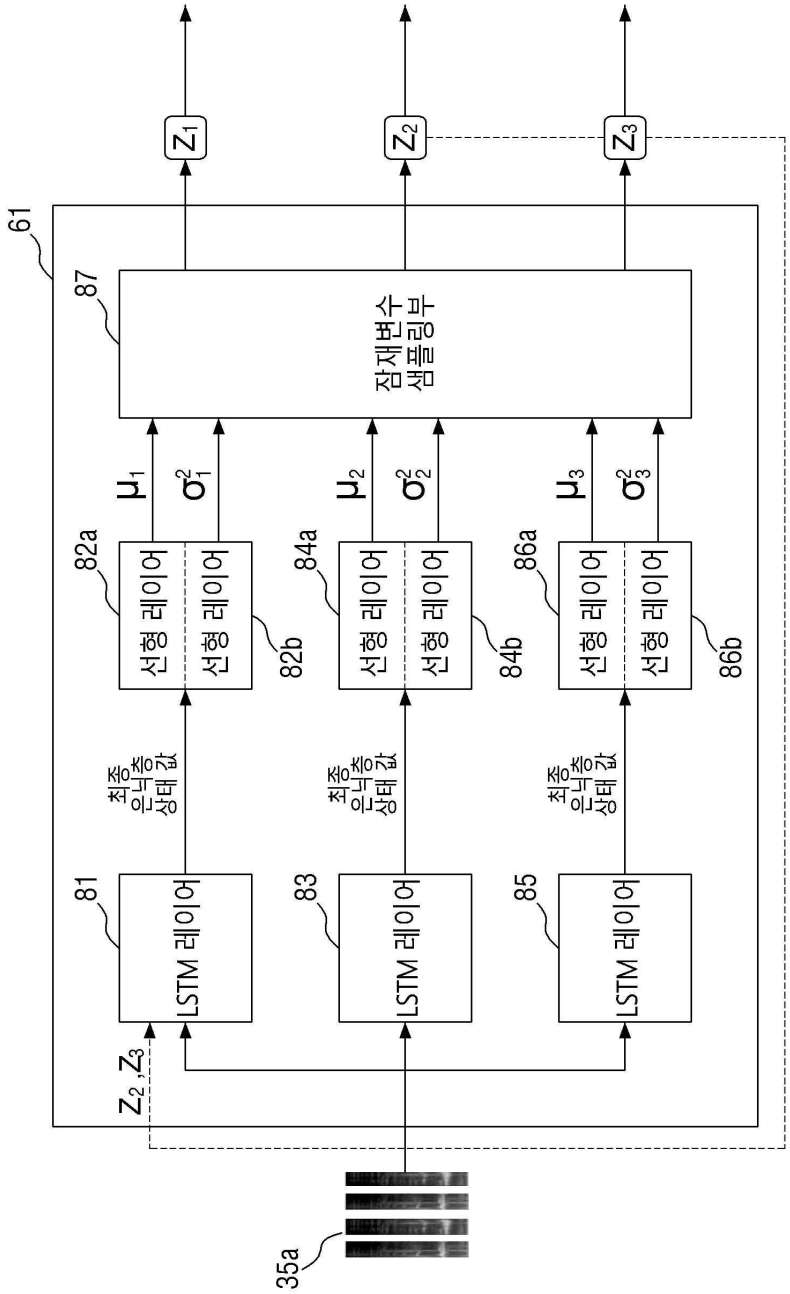
도면6


41a
41
53
69
디코딩부
$\tilde{z}_3$
67
잠재변수 변환부
z_1
z_2
z_3
감정 임베딩 메트릭스
63
61
인코딩부
35a
35

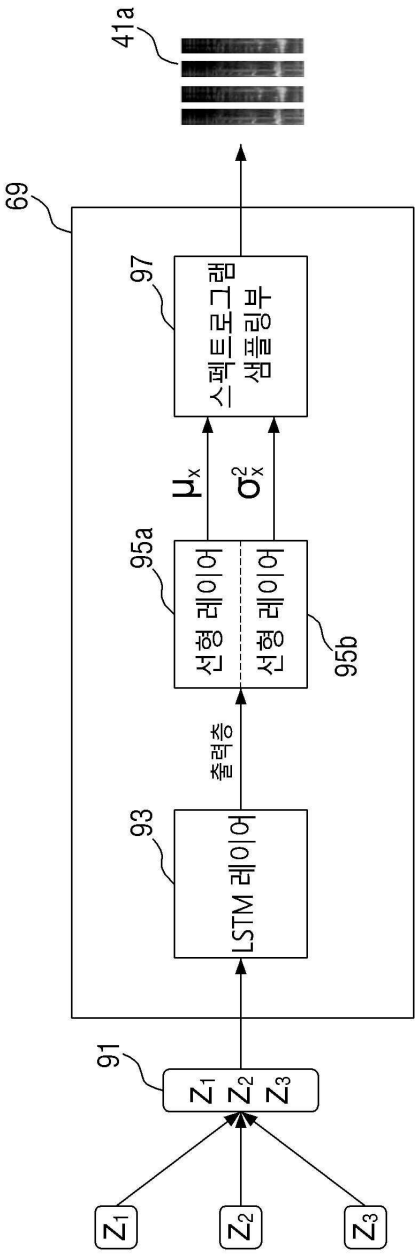
도면7



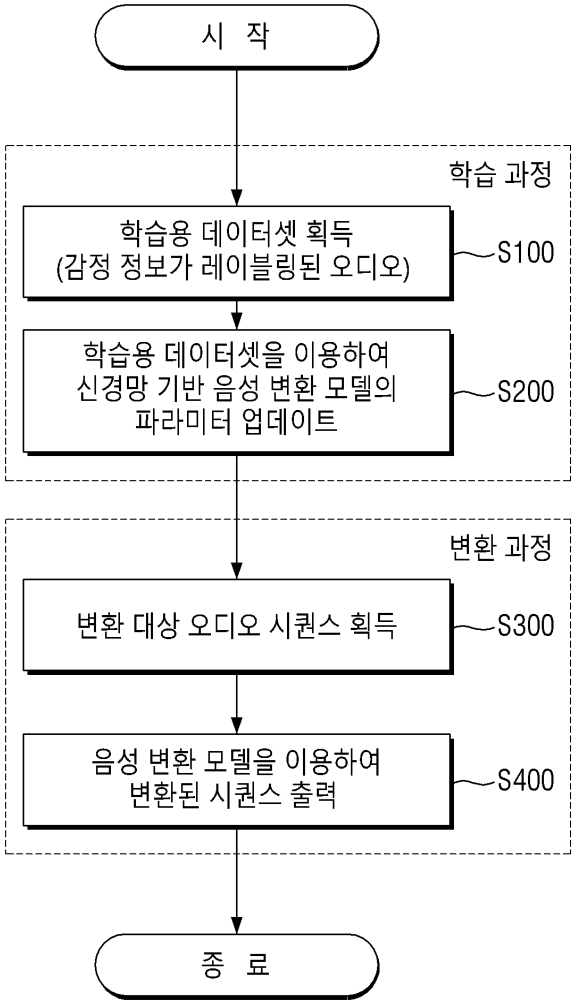
도면8


61
87
잠재변수 샘플링부
Z_1
Z_2
Z_3
μ_1
σ_1^2
μ_2
σ_2^2
μ_3
σ_3^2
82a
선형 레이어
82b
선형 레이어
84a
선형 레이어
84b
선형 레이어
86a
선형 레이어
86b
선형 레이어
최종 은닉층 상태 값
최종 은닉층 상태 값
최종 은닉층 상태 값
81
LSTM 레이어
83
LSTM 레이어
85
LSTM 레이어
Z_2, Z_3
35a

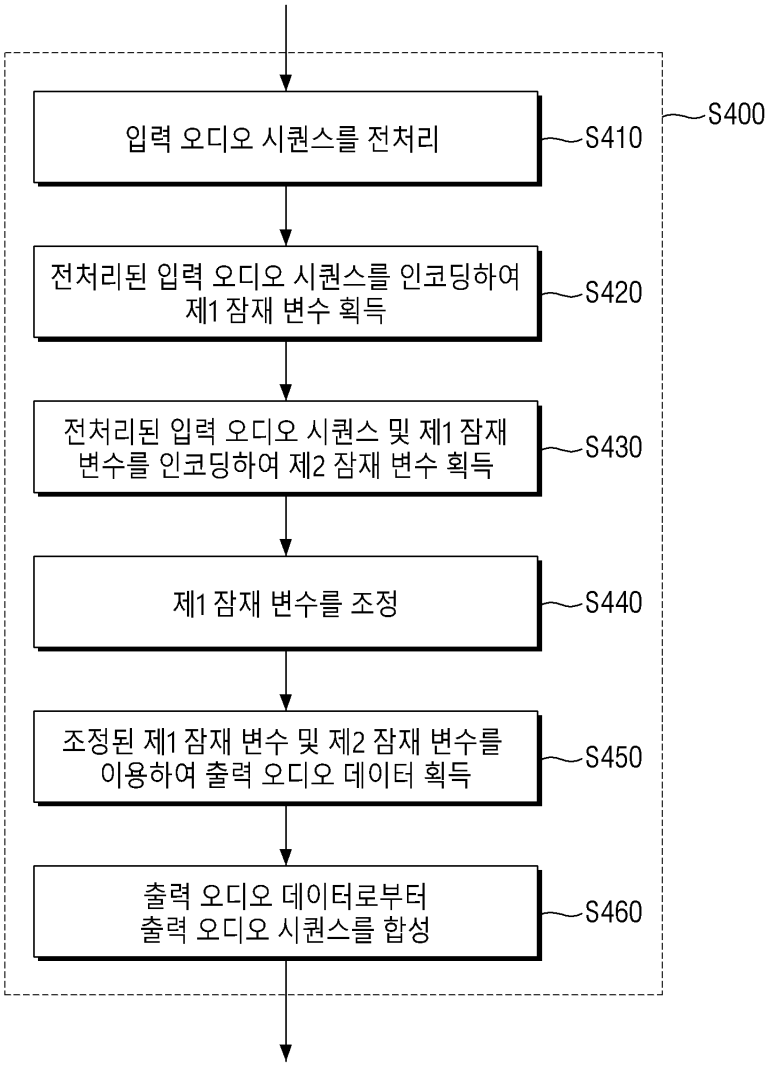
도면9


41a
69
97
스펙트로그램 샘플링부
μ_x
σ_x^2
95a
선형 레이어
선형 레이어
95b
출력층
93
LSTM 레이어
91
Z_1 Z_2 Z_3
Z_1
Z_2
Z_3

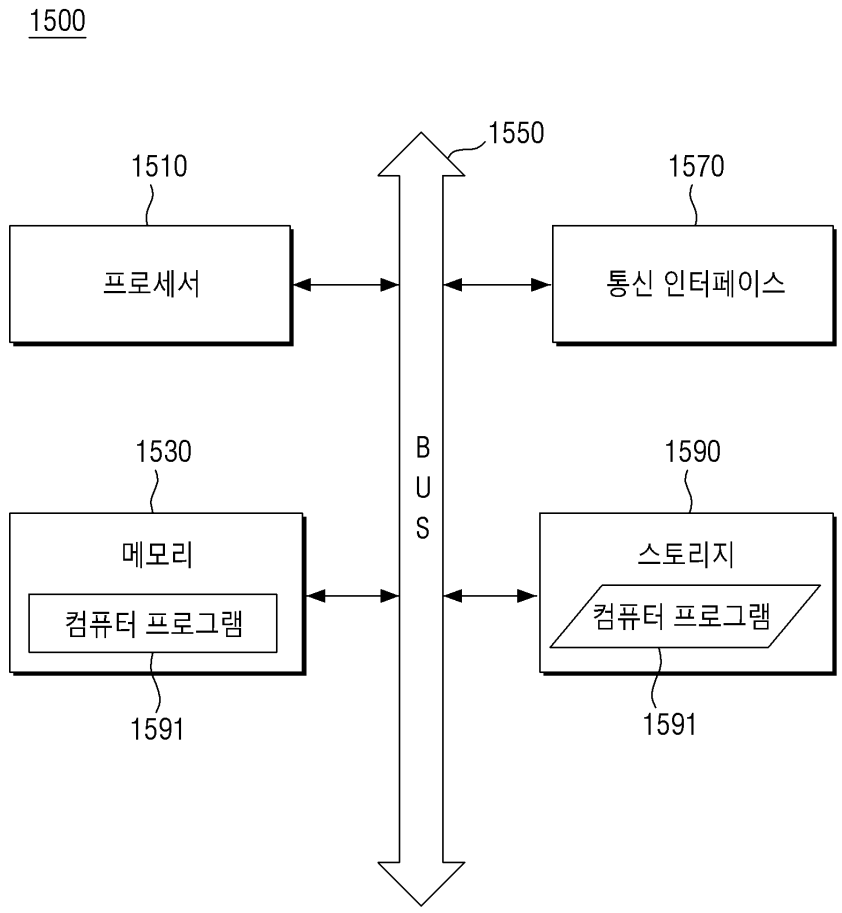
도면10


시 작
학습 과정
학습용 데이터셋 획득
(감정 정보가 레이블링된 오디오)
S100
학습용 데이터셋을 이용하여
신경망 기반 음성 변환 모델의
파라미터 업데이트
S200
변환 과정
변환 대상 오디오 시퀀스 획득
S300
음성 변환 모델을 이용하여
변환된 시퀀스 출력
S400
종 료

도면11


입력 오디오 시퀀스를 전처리
S410
S400
전처리된 입력 오디오 시퀀스를 인코딩하여 제1 잠재 변수 획득
S420
전처리된 입력 오디오 시퀀스 및 제1 잠재 변수를 인코딩하여 제2 잠재 변수 획득
S430
제1 잠재 변수를 조정
S440
조정된 제1 잠재 변수 및 제2 잠재 변수를 이용하여 출력 오디오 데이터 획득
S450
출력 오디오 데이터로부터 출력 오디오 시퀀스를 합성
S460

도면12


1500
1550
1510
1570
프로세서
통신 인터페이스
1530
1590
B
U
S
메모리
컴퓨터 프로그램
스토리지
컴퓨터 프로그램
1591
1591