(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0044337
(43) 공개일자 2020년04월29일

(51) 국제특허분류(Int. Cl.)
G10L 13/04 (2006.01) *G10L 13/08* (2006.01)
G10L 15/06 (2006.01) *G10L 25/63* (2013.01)
(52) CPC특허분류
G10L 13/043 (2013.01)
G10L 13/08 (2013.01)
(21) 출원번호 10-2018-0124925
(22) 출원일자 2018년10월19일
심사청구일자 없음

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
장인선
대전광역시 유성구 배울2로 78, 607동 1101호(관평동, 대덕테크노밸리6단지아파트)
안충현
대전광역시 유성구 대덕대로541번길 68, 103동 306호(도룡동, 대덕연구원현대아파트)
(뒷면에 계속)
(74) 대리인
성병기

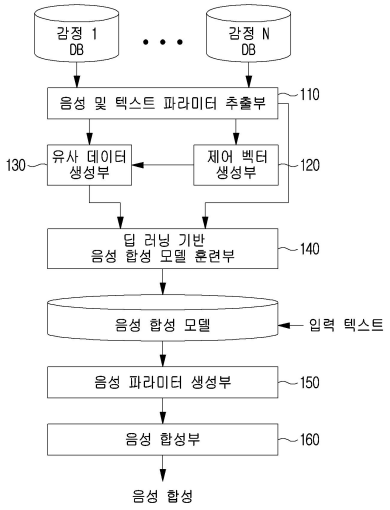
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **생성 모델 기반 데이터 증강 기법을 활용한 딥러닝 기반 감정음성합성 장치 및 방법**

(57) 요 약

본 발명은 음성 합성을 수행하는 방법 및 장치에 대한 것으로, 보다 상세하게는 유사 증강 데이터를 생성하여 음성합성 모델을 훈련하고, 유사 증강 데이터를 생성하는 경우 유사데이터 생성모델(generative model)에 상기 감정 조절 벡터를 입력하여 유사 증강 데이터를 생성하는 것을 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G10L 15/063 (2013.01)
G10L 25/63 (2013.01)
(72) 발명자
서정일
대전광역시 유성구 반석서로 109, 704동 1704호(반석동, 반석마을7단지아파트)
양승준
대전광역시 유성구 지족로 240, 504동 505호(지족동, 노은해랑숲마을5단지아파트)
최지훈
대전광역시 유성구 은구비남로33번길 56, 903동 1902호(지족동, 열매마을9단지)
강홍구
서울특별시 서대문구 연희로32길 48, 108동 702호(연희동, 연희동성원아파트)
강현주
서울특별시 서대문구 연세로 50
권오성
서울특별시 동대문구 정릉천동로 36, 107동 604호(용두동, 래미안허브리츠)

이 발명을 지원한 국가연구개발사업
과제고유번호 2015-0-00860
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터(IITP)
연구사업명 ETRI연구개발지원사업
연구과제명 시청각장애인 방송접근권 향상을 위한 디지털자막·음성해설 서비스 기술 개발
기 여 율 1/1
주관기관 ETRI
연구기간 2018.03.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

음성 합성 장치가 음성 합성을 수행하는 방법에 있어서

데이터 베이스로부터 음성데이터들을 입력 받고, 상기 음성데이터들로부터 파라미터들을 추출하는 단계;

상기 음성데이터들을 기초로 유사 증강 데이터들을 생성하는 단계;

상기 음성데이터들과 상기 유사 증강 데이터들에 기초하여 음성합성 모델을 훈련하는 단계; 및

텍스트를 입력 받고, 상기 음성합성 모델을 사용하여 음성을 합성하여 출력하는 단계; 를 포함하되,

상기 유사 증강 데이터를 생성하는 경우

감정 조절 벡터를 생성하고,

유사데이터 생성모델(generative model)에 상기 감정 조절 벡터를 입력하여,

적어도 하나 이상의 상기 유사 증강 데이터를 생성하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 2

제 1 항에 있어서

상기 감정 조절 벡터는

상기 유사데이터 생성모델의 입력으로 사용되는 확률 변수를 제어함으로써 감정 표현의 방법이나 감정 표현의 강도를 조절하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 3

제 1 항에 있어서

상기 생성된 적어도 하나 이상의 데이터는

상기 음성데이터와 유사한 확률 분포를 갖는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 4

제 1 항에 있어서

상기 유사데이터 생성모델은

상기 음성데이터들의 파라미터들을 입력 받아 훈련되는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 5

제 1 항에 있어서

상기 유사데이터 생성모델은

상기 음성데이터들로부터 추출한 파라미터들의 확률 분포를 통계적으로 모델링하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 6

제 1 항에 있어서

상기 유사데이터 생성모델을 훈련하는 경우,

VAE (Variational Auto Encoder)를 이용하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 7

제 1 항에 있어서

상기 음성데이터의 파라미터를 입력 받아 감정 조절 모델이 훈련되고,

상기 훈련된 감정 조절 모델은 랜덤 변수를 입력 받아 상기 감정 조절 벡터를 생성하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 8

제 1 항에 있어서

상기 음성합성 모델을 훈련하는 경우,

상기 음성데이터들로부터 추출된 파라미터들 및 유사 증강데이터들로부터 추출된 파라미터들에 기초하여 상기 음성합성 모델을 훈련하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 9

제 1 항에 있어서

음성합성 모델을 훈련하는 경우,

언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보를 저장하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 10

제 9 항에 있어서

상기 텍스트를 입력 받은 경우,

상기 음성합성 모델에 저장된 언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보에 기초하여,

상기 텍스트에 대응되는 음성 파라미터를 추정하여 음성 파형을 합성하여 출력하는 것을 특징으로 하는 음성 합성을 수행하는 방법.

청구항 11

음성합성 장치에 있어서

데이터 베이스로부터 음성데이터들을 입력 받고, 상기 음성데이터들로부터 파라미터들을 추출하는 파라미터 추출부;

상기 음성데이터들을 기초로 유사 증강 데이터들을 생성하는 유사 데이터 생성부;

상기 음성데이터들과 상기 유사 증강 데이터들에 기초하여 음성합성 모델을 훈련하는 음성합성 모델 훈련부; 및

텍스트를 입력 받고, 상기 음성합성 모델을 사용하여 음성을 합성하여 출력하는 음성 합성부; 를 포함하되,

상기 유사 데이터 생성부가 상기 유사 증강 데이터를 생성하는 경우

감정 조절 벡터를 생성하고,

유사데이터 생성모델(generative model)에 상기 감정 조절 벡터를 입력하여,

적어도 하나 이상의 데이터를 추가로 생성하는 것을 특징으로 하는 음성합성 장치.

청구항 12

제 11 항에 있어서

상기 감정 조절 벡터는

상기 유사데이터 생성모델의 입력으로 사용되는 확률 변수를 제어함으로써 감정 표현의 방법이나 감정 표현의 강도를 조절하는 것을 특징으로 하는 음성합성 장치.

청구항 13

제 11 항에 있어서

상기 생성된 적어도 하나 이상의 데이터는

상기 음성데이터와 유사한 확률 분포를 갖는 것을 특징으로 하는 음성합성 장치.

청구항 14

제 11 항에 있어서

상기 유사데이터 생성모델은

상기 음성데이터들의 파라미터들을 입력 받아 훈련되는 것을 특징으로 하는 음성합성 장치.

청구항 15

제 11 항에 있어서

상기 유사데이터 생성모델은

상기 음성데이터들로부터 추출한 파라미터들의 확률 분포를 통계적으로 모델링하는 것을 특징으로 하는 음성합성 장치.

청구항 16

제 11 항에 있어서

상기 유사데이터 생성모델을 훈련하는 경우,

VAE (Variational Auto Encoder)를 이용하는 것을 특징으로 하는 음성합성 장치.

청구항 17

제 1 항에 있어서

제어 벡터 생성부;를 더 포함하되,

상기 음성데이터의 파라미터를 입력 받아 감정 조절 모델이 훈련되고,

상기 훈련된 감정 조절 모델은 랜덤 변수를 입력 받아 상기 감정 조절 벡터를 생성하는 것을 특징으로 하는 음성합성 장치.

청구항 18

제 1 항에 있어서

상기 음성합성 모델을 훈련하는 경우,

상기 음성데이터들로부터 추출된 파라미터들 및 유사 증강데이터들로부터 추출된 파라미터들에 기초하여 상기 음성합성 모델을 훈련하는 것을 특징으로 하는 음성합성 장치.

청구항 19

제 11 항에 있어서

음성합성 모델을 훈련하는 경우,

언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보를 저장하는 것을 특징으로 하는 음성합성 장치.

청구항 20

제 19 항에 있어서

음성 파라미터 생성부;를 더 포함하고,

상기 텍스트를 입력 받은 경우,

상기 음성합성 모델에 저장된 언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보에 기초하여,

상기 음성 파라미터 생성부는 상기 텍스트에 대응되는 음성 파라미터를 추정하고,

음성 합성부는 상기 추정된 음성 파라미터에 기초하여 음성 파형을 합성하여 출력하는 것을 특징으로 하는 음성합성 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 음성 합성 시스템에 관한 것으로서, 보다 상세하게는 감정 음성 합성 시스템으로부터 고품질의 감정 음성을 합성하기 위한 효과적인 훈련 방법에 관한 것이다.

배 경 기 술

[0002] 음성 합성 시스템은 입력 텍스트를 사람이 직접 발화한 음성과 같이 청각적으로 자연스러운 음성 신호로 변환하여 출력하는 시스템을 가리킨다.

[0003] 통계적 파라메트릭 모델(statistical parametric model) 기반의 음성 합성 기법은 컨텍스트 정보에 따라 음성 파라미터를 모아 통계적으로 모델링한 후, 이를 활용하여 시스템에 입력된 텍스트에 대응하는 음성 파라미터를 생성하여 음성 신호를 합성한다. 이 방식은 수 시간 이내의 데이터베이스를 사용해도 양질의 합성음을 제공하며, 쉽게 음성 파라미터를 제어할 수 있고 이에 따라 합성음의 음색 등을 조절하기 용이하다는 장점 때문에 실생활에서 다양한 분야에 사용되고 있다. 이때 통계적 모델로는 입출력 데이터간의 비선형적이고 복잡한 관

계를 모델링할 수 있는 딥 러닝 모델이 널리 사용되고 있다.

[0004] 통계적 파라메트릭 모델 기반의 시스템에서 양호한 음질을 얻기 위해서는 일반적으로 최소 3시간 가량의 음성 데이터가 필요하지만, 감정 음성 합성 시스템에 사용되는 음성 데이터베이스는, 한 화자의 데이터베이스만을 제작하더라도 사람의 감정 종류는 다양하기 때문에 낭독체 음성 데이터베이스를 만들 때 보다 수 시간 이상의 시간과 비용을 필요로 한다. 또한 같은 감정을 표현하더라도 화자에 따라 표현 방식의 편차가 크기 때문에, 단순히 여러 화자의 데이터를 모아서 음성 합성기를 훈련하는 것은 화자 별 감정 표현의 차이들을 나타내기가 어려우므로 생동감 있는 감정 음성을 합성하지 못한다.

[0005] 따라서 감정 음성 데이터베이스의 경우, 감정의 종류가 매우 다양하기 때문에 데이터베이스를 구축하는 비용적 한계가 매우 큰 점 및 같은 감정을 표현하더라도 표현 방식에 있어 화자에 따른 편차가 큰 점을 고려하여 고품질의 합성음을 얻기 위한 방안이 필요한 실정이다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 감정 음성 합성 시스템에서 합성음의 품질을 향상시키기 위한 목적이 있다.

[0007] 본 발명은 딥 러닝 기반의 감정 음성 합성 시스템을 효과적으로 훈련하여 고품질의 합성 음성을 제공하는데 목적이 있다.

[0008] 본 발명은 딥 러닝 기반의 감정 음성 합성 시스템을 효과적으로 훈련하기 위하여 데이터를 증강하여 생성하는데 목적이 있다.

[0009] 본 발명에서 이루고자 하는 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급하지 않은 또 다른 기술적 과제들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 본 발명의 일 실시예에 따라, 음성 합성 장치가 음성 합성을 수행하는 방법 및 장치를 제공할 수 있다. 이 때 음성 합성 장치는 파라미터 추출부, 유사 데이터 생성부, 음성합성 모델 훈련부 및 음성 합성부를 포함할 수 있다.

[0011] 이 때, 파라미터 추출부는 데이터 베이스로부터 음성데이터들을 입력 받고, 상기 음성데이터들로부터 파라미터들을 추출할 수 있다.

[0012] 유사 데이터 생성부는 음성데이터들을 기초로 유사 증강 데이터들을 생성할 수 있다.

[0013] 음성합성 모델 훈련부는 음성데이터들과 유사 증강 데이터들에 기초하여 음성합성 모델을 훈련할 수 있다.

[0014] 음성 합성부는 텍스트를 입력 받고, 음성합성 모델을 사용하여 음성을 합성하여 출력할 수 있다.

[0015] 이 때 상기 유사 증강 데이터를 생성하는 경우, 감정 조절 벡터를 생성하고, 유사데이터 생성모델(generative model)에 상기 감정 조절 벡터를 입력하여, 적어도 하나 이상의 상기 유사 증강 데이터를 생성할 수 있다.

[0016] 또한, 다음의 실시예들은 음성 합성 장치가 음성 합성을 수행하는 방법 및 장치에서 공통으로 적용될 수 있다.

[0017] 본 발명의 일 실시예에 따라, 감정 조절 벡터는 유사데이터 생성모델의 입력으로 사용되는 확률 변수를 제어함으로써 감정 표현의 방법이나 감정 표현의 강도를 조절할 수 있다.

[0018] 본 발명의 일 실시예에 따라, 생성된 적어도 하나 이상의 데이터는 음성데이터와 유사한 확률 분포를 가질 수 있다.

[0019] 본 발명의 일 실시예에 따라, 유사데이터 생성모델은 음성데이터들의 파라미터들을 입력 받아 훈련될 수 있다.

[0020] 본 발명의 일 실시예에 따라, 유사데이터 생성모델은 음성데이터들로부터 추출한 파라미터들의 확률 분포를 통계적으로 모델링할 수 있다.

[0021] 본 발명의 일 실시예에 따라, 유사데이터 생성모델을 훈련하는 경우, VAE (Variational Auto Encoder)를 이용할 수 있다.

[0022] 본 발명의 일 실시예에 따라, 음성데이터의 파라미터를 입력 받아 감정 조절 모델이 훈련되고, 훈련된 감정 조절 모델은 랜덤 변수를 입력 받아 감정 조절 벡터를 생성할 수 있다.

[0023] 본 발명의 일 실시예에 따라, 음성합성 모델을 훈련하는 경우, 음성데이터들로부터 추출된 파라미터들 및 유사 증강데이터들로부터 추출된 파라미터들에 기초하여 상기 음성합성 모델을 훈련할 수 있다.

[0024] 본 발명의 일 실시예에 따라, 음성합성 모델을 훈련하는 경우, 언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보를 저장할 수 있다.

[0025] 본 발명의 일 실시예에 따라, 텍스트를 입력 받은 경우, 음성합성 모델에 저장된 언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보에 기초하여, 텍스트에 대응되는 음성 파라미터를 추정하여 음성 파형을 합성하여 출력할 수 있다.

발명의 효과

[0026] 본 발명에 의하면 감정 음성 합성 시스템에서 합성음의 품질을 향상시킬 수 있다.

[0027] 본 발명에 의하면 딥 러닝 기반의 감정 음성 합성 시스템을 효과적으로 훈련하여 고품질의 합성 음성을 제공할 수 있다.

[0028] 본 발명에 의하면 딥 러닝 기반의 감정 음성 합성 시스템을 효과적으로 훈련하기 위하여 데이터를 증강하여 생성할 수 있다.

[0029] 본 발명에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0030] 도 1은 본 발명의 일 실시예에 따른 음성 합성 장치의 구성을 나타낸 도면이다.

도 2는 본 발명의 실시예에 따른 음성 합성 방법의 흐름도이다.

도 3은 본 발명의 실시예에 따른 음성 합성 장치의 구성 및 구체적인 흐름도이다.

도 4는 본 발명의 실시예에 따른 증강 데이터를 생성하는 방법의 흐름도이다.

도 5는 제어벡터부의 훈련 및 벡터 생성 과정에 대한 도면이다.

도 6은 유사 데이터 생성부의 동작에 대한 흐름도이다.

도 7은 딥 러닝 기반 음성 합성기의 전체 개요도 이다.

발명을 실시하기 위한 구체적인 내용

[0031] 이하에서는 첨부한 도면을 참고로 하여 본 발명의 실시 예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다.

[0032] 본 발명의 실시 예를 설명함에 있어서 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그에 대한 상세한 설명은 생략한다. 그리고, 도면에서 본 발명에 대한 설명과 관계없는 부분은 생략하였으며, 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0033] 본 발명에 있어서, 어떤 구성요소가 다른 구성요소와 "연결", "결합" 또는 "접속"되어 있다고 할 때, 이는 직접적인 연결관계뿐만 아니라, 그 중간에 또 다른 구성요소가 존재하는 간접적인 연결관계도 포함할 수 있다. 또한 어떤 구성요소가 다른 구성요소를 "포함한다" 또는 "가진다"고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 배제하는 것이 아니라 또 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0034] 본 발명에 있어서, 제1, 제2 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용되며, 특별히 언급되지 않는 한 구성요소들간의 순서 또는 중요도 등을 한정하지 않는다. 따라서, 본 발명의 범위 내에서 일 실시 예에서의 제1 구성요소는 다른 실시 예에서 제2 구성요소라고 칭할 수도 있고, 마찬가지로 일

실시 예에서의 제2 구성요소를 다른 실시 예에서 제1 구성요소라고 칭할 수도 있다.

[0035] 본 발명에 있어서, 서로 구별되는 구성요소들은 각각의 특징을 명확하게 설명하기 위함이며, 구성요소들이 반드시 분리되는 것을 의미하지는 않는다. 즉, 복수의 구성요소가 통합되어 하나의 하드웨어 또는 소프트웨어 단위로 이루어질 수도 있고, 하나의 구성요소가 분산되어 복수의 하드웨어 또는 소프트웨어 단위로 이루어질 수도 있다. 따라서, 별도로 언급하지 않더라도 이와 같이 통합된 또는 분산된 실시 예도 본 발명의 범위에 포함된다.

[0036] 본 발명에 있어서, 다양한 실시 예에서 설명하는 구성요소들이 반드시 필수적인 구성요소들은 의미하는 것은 아니며, 일부는 선택적인 구성요소일 수 있다. 따라서, 일 실시 예에서 설명하는 구성요소들의 부분집합으로 구성되는 실시예도 본 발명의 범위에 포함된다. 또한, 다양한 실시 예에서 설명하는 구성요소들에 추가적으로 다른 구성요소를 포함하는 실시 예도 본 발명의 범위에 포함된다.

[0037] 딥 러닝 기반 감정 음성 합성 시스템에서 감정을 생생하게 표현하는 고품질의 합성음을 얻기 위해서는 충분한 양의 음성 데이터를 필요로 한다. 그러나 실제 상황에서 대용량의 감정 음성 데이터를 구축하는 것은 많은 비용과 시간이 소모되므로 어려운 일이다.

[0038] 언어적 정보의 경우 한 개의 문자에 대해 한 개의 발음 정보가 대응하므로 일대일 대응 관계이다. 따라서 다양한 사람들이 발화하더라도 동일한 문자의 경우 발음 정보에 대해 공통적인 특성이 뚜렷하다. 그러나 감정 음성을 발화할 경우, 같은 감정을 표현하더라도 표현 강도 및 방법에 있어 화자마다 편차가 매우 크므로 이는 일대다 대응 관계에 가깝다.

[0039] 따라서, 본 발명은 딥 러닝 기반의 통계적 생성 모델링 기법을 활용하여 실제 감정 음성 데이터베이스로부터 추출한 것과 유사한 통계적 특성을 가지는 파라미터를 생성하며, 이 때 감정의 강도 혹은 감정의 표현 방식 등을 제어함으로써 다양한 감정 표현과 강도의 조절을 가능하게 한다. 이렇게 생성한 데이터를 원래의 감정 음성 데이터베이스와 함께 딥 러닝 기반의 음성 합성 시스템을 훈련용 데이터로 사용하여 대용량 데이터베이스로 훈련한 음성 합성 시스템과 같은 결과를 얻음으로써 고품질의 합성음과 다양한 감정 표현이 가능한 감정 음성 합성 시스템을 제공하는 것을 목표로 한다.

[0040] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예를 상세히 설명하기로 한다.

[0041] 도1은 본 발명의 일 실시예에 따른 음성 합성 장치의 구성을 나타낸 도면이다.

[0042] 도 1을 참조하면, 본 발명의 음성 합성 시스템은 음성 파라미터 추출부(110), 제어 벡터 생성부(120), 유사 데이터 생성부(130), 딥 러닝 기반의 음성 합성 모델 훈련부(140), 음성 파라미터 생성부(150), 음성 합성부(160)를 포함한다.

[0043] 본 발명은 딥 러닝 기반의 생성 모델을 이용하여 실제 음성 신호로부터 추출한 것과 유사한 확률 분포를 갖는 파라미터를 감정 표현 혹은 감정 강도에 따라 확률 변수를 제어하여 필요에 맞게 생성하고, 이렇게 생성한 파라미터를 이용하여 딥 러닝 기반의 감정 음성 합성 시스템을 효과적으로 훈련하여 고품질의 합성 음성을 제공하는 데에 있다.

[0044] 본 발명에 따른 생성 모델 기반의 훈련 데이터(Training Data) 증강 기법을 활용한 딥 러닝 기반 감정 음성 합성 시스템의 효과적인 훈련 방법은, 감정 음성 합성을 위한 데이터를 사용하여 확률 분포생성 모델을 훈련하고, 이렇게 얻은 모델을 이용하여 실제 데이터의 확률 분포를 따르는 유사 데이터들을 생성하고, 이 때 미리 훈련해 둔 제어 벡터(본 발명의 일 실시예로, 감정 표현 방식, 감정 강도 등을 제어)를 이용하여 목적에 맞도록 감정 표현 특징을 제어할 수 있으며, 이렇게 제어 벡터를 이용하여 생성한 유사 데이터를 원 데이터베이스로부터 추출한 파라미터와 함께 딥 러닝 기반의 음성 합성기의 훈련에 사용한다.

[0045] 이에 따라, 음성 합성 모델을 훈련할 때 더욱 풍부한 데이터를 사용할 수 있으며, 이에 따라 모델을 더욱 세밀하게 조정하여 훈련할 수 있고, 해당 모델을 통해 동적 정보 (dynamic)가 보존된 파라미터를 생성함으로써 데이터의 정확도를 개선할 수 있으며, 더 나아가 다양한 감정 표현 방식과 표현 강도를 제어함으로써 다양한 서비스에 활용될 수 있다.

[0046] 따라서 본 기술은 이러한 문제점을 극복하기 위해 데이터의 확률 분포를 통계적으로 모델링하는 생성 모델(generative model)을 이용한다. 생성 모델을 이용하면 훈련 데이터(Training Data)와 유사한 데이터를 만들어 낼 수 있는데, 이를 통해 음성 합성에 사용되는 감정 음성 파라미터와 유사한 데이터를 생성하여 음성 데이터베이스의 양을 늘리는 것과 같은 효과를 얻고자 한다.

[0047] 이 때 생성 모델의 입력으로 사용되는 확률 변수를 제어함으로써 감정 표현의 방법이나 감정 표현의 강도를 조절할 수 있으며, 이를 통해 세밀하게 감정을 조절할 수 있고 대용량 데이터베이스를 이용하여 훈련한 것과 같은 고품질의 감정 음성을 합성할 수 있다.

[0048] 도 2는 본 발명의 실시예에 따른 음성 합성 방법의 흐름도이다.

[0049] 음성 합성을 수행하기 위해, 증강데이터를 생성하고(S210), 생성된 증강데이터를 이용하여 음성 합성기를 훈련하고(S220), 음성 합성을 출력(S230)한다.

[0050] 따라서, 본 장치를 크게 구분 하였을 때, 감정 음성 합성 시스템에 사용되는 음성 데이터베이스의 데이터를 증강시키기는 구성(110, 120, 130)과 증강된 데이터를 활용하여 음성합성 모델을 훈련하는 구성(140) 및 음성 합성을 출력하는 구성(150, 160)으로 구분할 수 있다.

[0051] 도 3은 본 발명의 실시예에 따른 음성 합성 장치의 구성 및 구체적인 흐름도이다.

[0052] 음성 파라미터 추출부(110)는 감정 음성 합성 데이터베이스로부터 언어 및 음성 파라미터를 추출한다.

[0053] 그리고 제어 벡터 생성부(120) 및 유사 데이터 생성부(130)은 유사 데이터들을 생성하여 음성 데이터 베이스의 양을 증가시킬 수 있다. 제어 벡터 생성부(120)의 구체적 동작은 하기의 도 5, 유사 데이터 생성부(130)의 구체적 동작은 하기의 도 6에서 각각 상세히 기술된다.

[0054] 최근, 데이터의 확률 분포를 딥 러닝 기법을 이용해 모델링함으로써 실제 데이터와 유사한 데이터를 샘플링할 수 있는 생성 모델 (generative model)이 각광받고 있다. 기존의 딥 러닝 기반의 훈련 방법은 지도 학습(supervised learning) 관점에서 입력 데이터와 해당 데이터의 레이블 정보 간의 매핑 관계를 나타내는 함수를 학습하는 경우에 주로 적용되어 왔으나, 생성 모델의 경우 비지도 학습 (unsupervised learning) 관점에서 레이블 없이 주어진 데이터만을 이용하여 해당 데이터에 내재되어 있는 구조를 학습하는 것을 목표로 한다. 생성 모델을 이용할 경우 주어진 데이터(원 데이터)와 같은 확률 분포를 갖는 새로운 샘플(유사 데이터)을 생성해낼 수 있으며, 생성한 데이터를 실제 제작한 데이터베이스처럼 음성 합성 시스템의 통계적 훈련에 사용할 수 있다. 또한 생성 모델을 통해 데이터의 잠재 변수를 모델링할 수 있으며, 이를 이용하여 유의미한 특성들이 제어된 데이터를 생성할 수 있다.

[0055] 딥 러닝 기반의 음성 합성 모델 훈련부(140)는 상기 언급된 원 데이터베이스로부터 추출한 언어 및 음성 파라미터와 유사 데이터 생성부로부터 취득한 파라미터를 이용하여 함께 딥 러닝 기반의 음성 합성 모델을 훈련하여 언어 및 음성 파라미터 간의 매핑 (mapping) 관계에 대한 정보를 저장한다.

[0056] 음성 파라미터 생성부(150)는 음성 합성 모델에 저장된 매핑 정보를 이용하여 입력으로 주어진 텍스트에 대응하는 음성 파라미터를 추정한다.

[0057] 음성 합성부(160)는 상기 언급한 음성 파라미터 생성부로부터 추정한 음성 파라미터를 이용하여 음성 파형을 합성하여 출력한다.

[0058] 도 4는 본 발명의 실시예에 따른 증강 데이터를 생성하는 방법의 흐름도이다.

[0059] 생성 모델 기반의 훈련 데이터(Training Data) 증강 기법을 활용한 딥 러닝 기반 감정 음성 합성 장치의 효과적인 훈련 방법을 실시하기 위하여, 먼저 파라미터 추출부(110)는 데이터 베이스로부터 음성데이터들을 입력받고, 음성데이터들로부터 파라미터들을 추출(S410)한다.

[0060] 그리고 유사 데이터 생성부(130)는 음성데이터들을 기초로 유사 증강 데이터들을 생성한다.

[0061] 보다 상세하게는 먼저 유사 데이터 생성부(130)는 감정 음성 파라미터를 이용하여 유사 데이터 생성을 위한 통계적 모델을 훈련하고, 유사 데이터 모델을 생성(S420)한다.

[0062] 그리고 제어 벡터 생성부(120)은 감정 표현을 위한 잠재 변수 모델을 훈련한다. 즉, 감정을 나타내는 제어 변수를 생성하기 위한 통계적 모델을 훈련하고, 제어 변수 모델을 생성(S430)한다.

[0063] 그 후 제어 벡터 생성부(120)는 감정 잠재 변수 모델을 이용하여 감정 조절 확률 벡터를 생성(S440)한다. 즉, 제어 벡터 생성부(120)는 생성한 제어 변수 모델을 이용하여 목적에 맞는 제어 변수를 추정할 수 있다.

[0064] 그리고 유사 데이터 생성부(130)는 제어 벡터 생성부(120)에서 생성된 감정 조절 확률 벡터를 이용하여 감정 특성이 제어된 증강 데이터를 생성(S450)한다. 보다 상세하게는 추정한 제어변수를 입력으로 사용하여 유사 데이

터 생성 모델로부터 음성 합성 시스템의 훈련에 사용되는 유사 데이터를 생성할 수 있다.

[0065] 그 후 음성합성 모델 훈련부(140)는 상기 음성데이터들과 상기 유사 증강 데이터들에 기초하여 딥 러닝 기반의 음성 합성기를 훈련하고, 음성 합성부(160)는 텍스트를 입력 받고, 음성합성 모델을 사용하여 음성을 합성하여 출력한다.

[0066] 도 5는 제어벡터부의 훈련 및 벡터 생성 과정에 대한 도면이다. 보다 상세하게는 다양한 감정 표현 혹은 감정의 강도 조절을 가능하게 하는 제어 벡터부(120)의 훈련 및 벡터 생성 과정을 보다 자세하게 설명하기 위한 블록 다이어그램이다.

[0067] 제어 벡터 생성부(120)는 먼저 각 감정 음성 데이터베이스로부터 음성 파라미터를 추출한다. 그 후 제어 벡터 생성부(120)는 추출한 음성 파라미터를 이용하여 감정 음성으로부터 감정의 특성을 표현할 수 있는 잠재 변수를 모델링한다. 이를 통해 제어 벡터 생성부(120)는 유사 데이터 생성부(130)에서 생성하고자 하는 데이터의 조건을 표현할 수 있는 잠재 변수를 생성할 수 있다.

[0068] 보다 상세하게는 제어 벡터 생성부(120)는 각 감정 음성 데이터 베이스로부터 음성 파라미터를 추출한다. 그리고 제어 벡터 생성부(120)는 감정 조절 모델을 훈련하고, 감정 조절 모델을 생성한다. 제어 벡터 생성부(120)는 랜덤 변수를 입력받아 감정 조절 벡터를 추정하고, 감정 조절 확률 벡터(감정 조절 벡터)를 생성한다.

[0069] 이때, 랜덤 변수는 사용자의 지정 값이며, 감정 조절 벡터를 추정하기 위한 입력값이 될 수 있다.

[0070] 도 6은 유사 데이터 생성부의 동작에 대한 흐름도이다. 보다 상세하게는 도 6은 실제 음성 합성 데이터베이스로부터 추출한 파라미터들과 유사한 통계적 특징을 가지는 파라미터들을 생성하여 음성 합성 시스템을 훈련할 때 대용량의 음성 데이터베이스를 사용하는 것과 같은 효과를 기대할 수 있는 유사 데이터 생성부(130)를 상세하게 서술하는 블록 다이어그램이다.

[0071] 도 6의 유사 데이터 생성부(130)는 각 감정 음성 데이터베이스로부터 음성 합성에 필요한 파라미터를 추출한다. 그 후 유사 데이터 생성부(130)는 해당 음성 파라미터의 확률 분포를 통계적으로 모델링하는 생성 모델을 훈련한다. 이 훈련 과정을 통해 해당 모델은 음성 파라미터에 내재되어 있는 각종 특징들을 압축적으로 표현할 수 있다. 이 모델을 이용하여 실제 음성 데이터로부터 추출한 파라미터와 유사한 특징을 보이지만, 실제로 음성 신호로부터 추출하지 않은 유사 데이터를 생성할 수 있다.

[0072] 일 실시 예로, 유사 데이터 생성부(130)는 VAE (Variational Auto Encoder) 등을 이용하여 생성 모델을 훈련하고 유사 데이터를 생성할 수 있다. VAE 구조는 생성 모델 구조 중 하나로, 원 데이터의 확률 분포를 구하는 것을 목적으로 하는데, 이 때 최대 우도 (maximum lilkelihood)를 기준으로 하여 최적화하는 것을 목표로 한다.

[0073] 기존의 생성 모델을 훈련할 때 잠재 변수의 형태가 미분이 불가능하여 역전파 (backpropagation) 기법을 사용할 수 없었고, 대신 sampling기법을 많이 사용하였으나 이는 계산량이 매우 커서 긴 훈련 시간을 필요로 했다. 그러나 variational autoencoder의 경우 이를 미분 가능한 형태로 변환하여 역전파 기법을 사용할 수 있게 되었다. 이 때 VAE의 구조는 주어진 데이터를 잠재 변수의 형태로 압축하고, 잠재 변수를 입력으로 사용했을 때 주어진 데이터와 같은 형태의 데이터를 생성해낼 수 있다. 이러한 구조는 Autoencoder 구조의 encoder-decoder 구조와 유사했기 때문에 variational autoencoder라는 명칭을 갖고 있다.

[0074] 도 7은 딥 러닝 기반 음성 합성기의 전체 개요도 이다.

[0075] 상술한 바와 같이, 본 발명은 더욱 풍부한 데이터를 감정 음성 합성기의 훈련에 사용할 수 있도록 딥 러닝 기반의 생성 모델을 통해 실제 데이터와 유사한 확률분포를 갖는 파라미터를 제어 변수를 사용하여 감정 표현 정보를 목적에 맞도록 변조하여 생성한다. 이를 통해 합성음의 감정 표현 강도나 감정 표현의 방식을 조절할 수 있다.

[0076] 본 발명의 이점 및 특징, 그것들을 달성하는 방법은 첨부되어 있는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 제시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

부호의 설명

[0077] 110: 파라미터 추출부

120: 제어 벡터 생성부

130: 유사 데이터 생성부

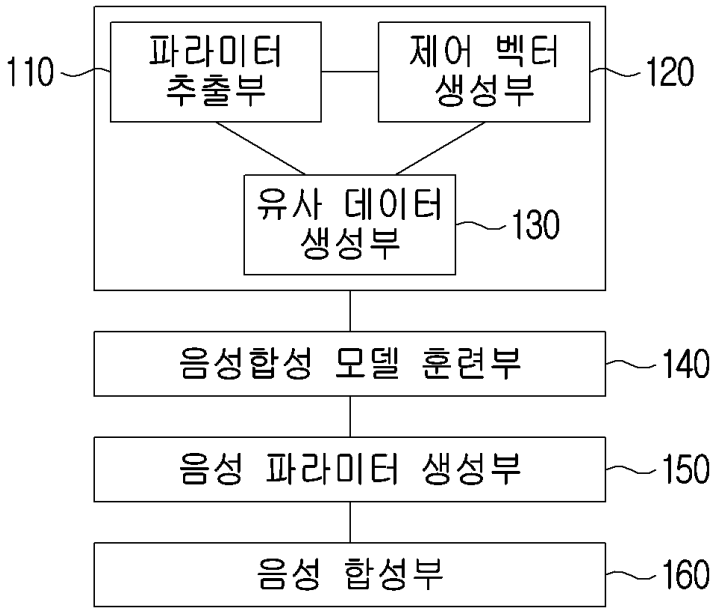
140: 음성합성 모델 훈련부

150: 음성 파라미터 생성부

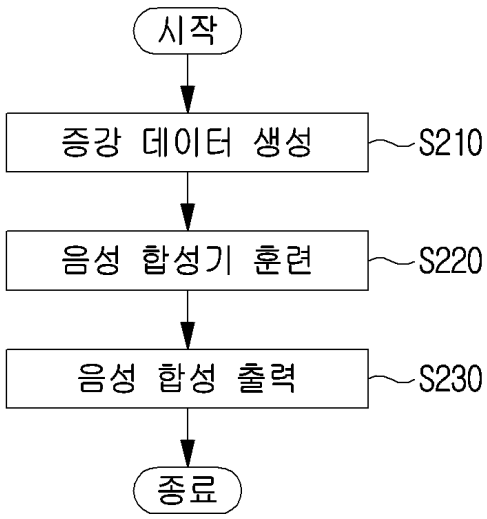
160: 음성 합성부

도면

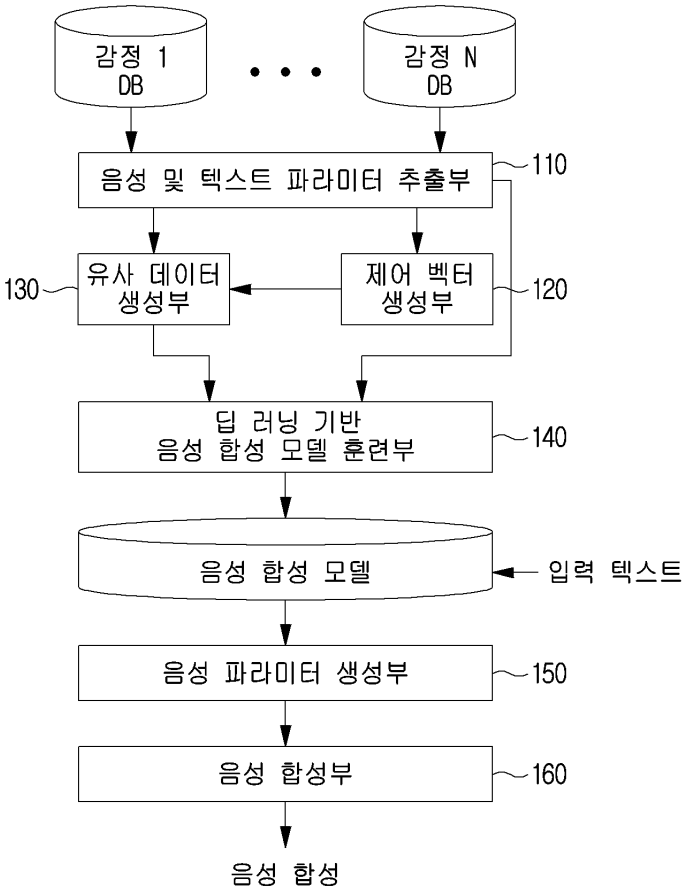
도면1



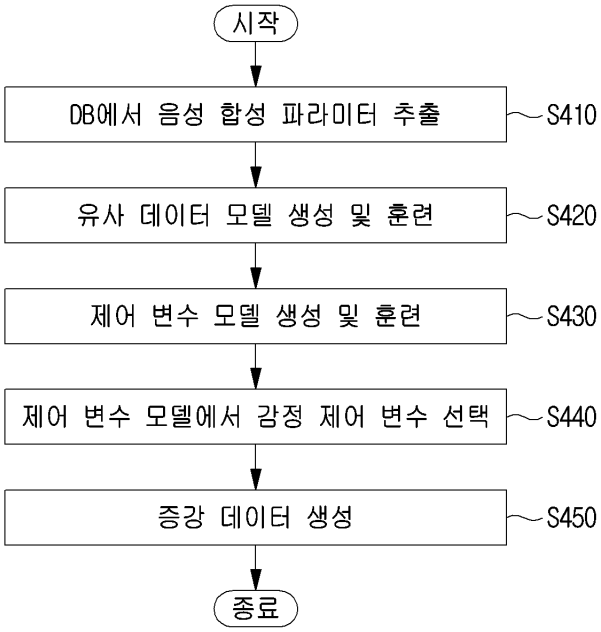
도면2



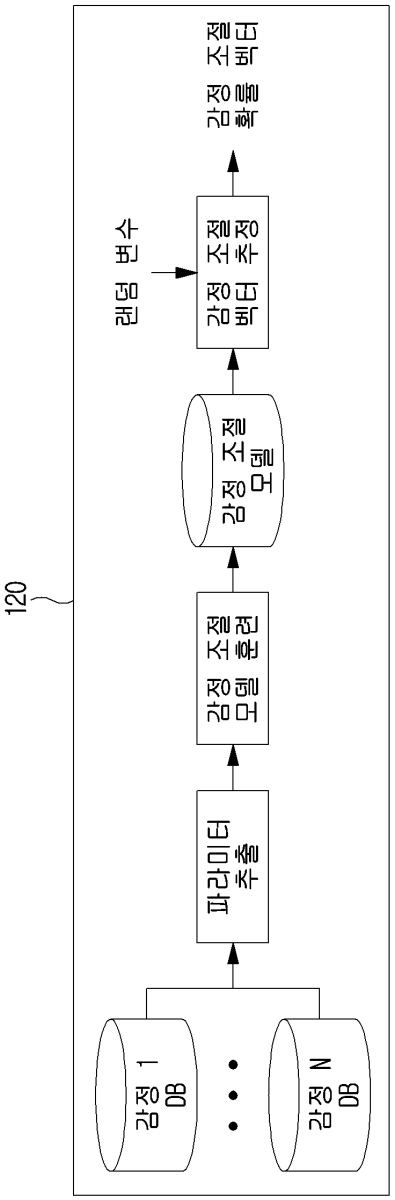
도면3


감정 1
DB
감정 N
DB
음성 및 텍스트 파라미터 추출부
110
유사 데이터
생성부
130
제어 벡터
생성부
120
딥 러닝 기반
음성 합성 모델 훈련부
140
음성 합성 모델
입력 텍스트
음성 파라미터 생성부
150
음성 합성부
160
음성 합성

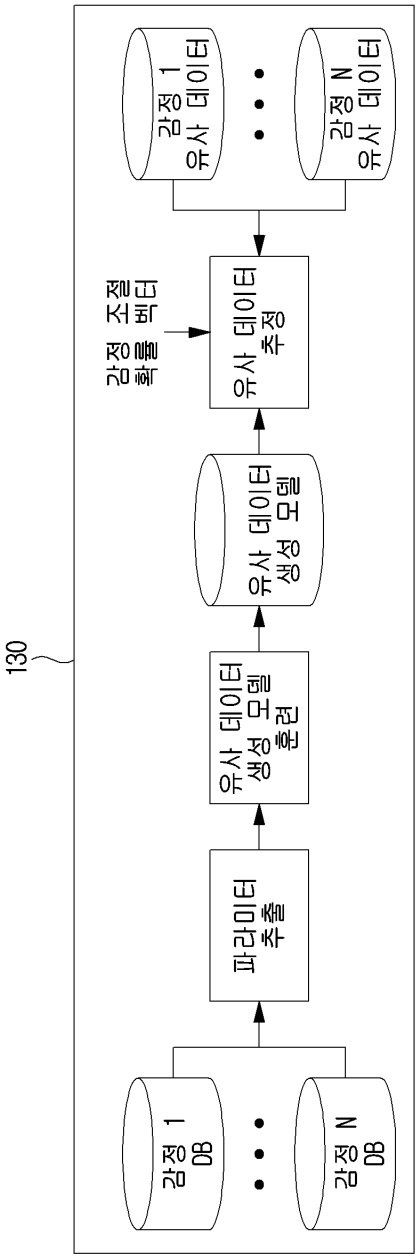
도면4


시작
DB에서 음성 합성 파라미터 추출
S410
유사 데이터 모델 생성 및 훈련
S420
제어 변수 모델 생성 및 훈련
S430
제어 변수 모델에서 감정 제어 변수 선택
S440
증강 데이터 생성
S450
종료

도면5


120
감정 1 DB
감정 N DB
파라미터 추출
감정 조절 모델 훈련
감정 조절 모델
랜덤 변수
감정 조절 벡터 추정
감정 조절 벡터 확률

도면6


130
감정 1 DB
감정 N DB
파라미터 추출
유사 데이터 생성 모델 훈련
유사 데이터 생성 모델
감정 조절 벡터
감정 확률
유사 데이터 추정
감정 1 유사 데이터
감정 N 유사 데이터

도면7