(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2025년01월03일
(11) 등록번호 10-2749818
(24) 등록일자 2024년12월30일

(51) 국제특허분류(Int. Cl.)
G10L 21/0208 (2013.01) *G10L 13/02* (2006.01)
G10L 25/18 (2013.01)
(52) CPC특허분류
G10L 21/0208 (2013.01)
G10L 13/02 (2013.01)
(21) 출원번호 10-2022-0175343
(22) 출원일자 2022년12월14일
심사청구일자 2022년12월14일
(65) 공개번호 10-2024-0092502
(43) 공개일자 2024년06월24일
(56) 선행기술조사문헌
Huajian Fang et al., 'VARIATIONAL AUTOENCODER FOR SPEECH ENHANCEMENT WITH A NOISE-AWARE ENCODER', ICASSP 2021, June 2021.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)
(72) 발명자
김남수
서울특별시 관악구 관악로 1, 서울대학교
김세민
서울특별시 관악구 청룡7길 29
(74) 대리인
김건우

전체 청구항 수 : 총 3 항 심사관 : 정성윤

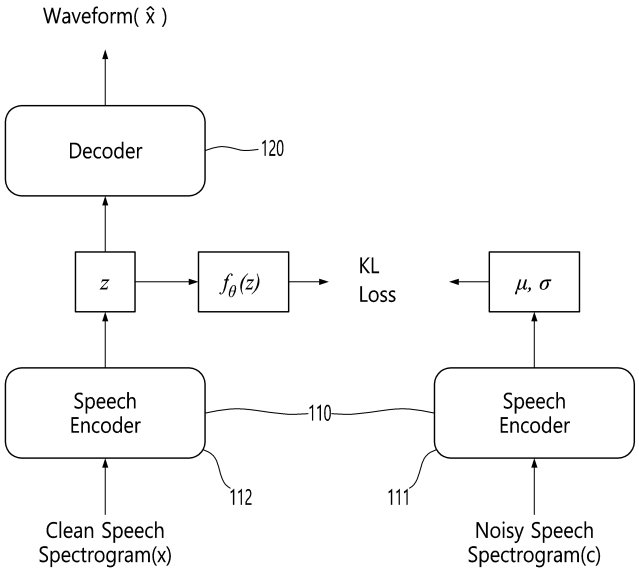
(54) 발명의 명칭 **음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체**

(57) 요 약

본 발명은 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 관한 것으로서, 보다 구체적으로는 음성 합성 구조 기반의 음성 향상 시스템으로서, 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각

(뒷면에 계속)

대 표 도 - 도4



의 잠재 변수(latent variable)로 인코딩 하는 음성 인코더(speech encoder); 및 상기 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더를 포함하는 것을 그 구성상의 특징으로 한다.

또한, 본 발명의 특징에 따른 음성 합성 구조 기반의 음성 향상 방법은, 음성 합성 구조 기반의 음성 향상 방법으로서, (1) 음성 인코더가 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 단계; (2) 디코더가 상기 단계 (1)의 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 단계를 포함하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 따르면, 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 음성 인코더와, 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더를 포함하여 구성함으로써, 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 할 수 있다.

또한, 본 발명의 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 따르면, 잡음이 있는 멜-스펙트로그램과 잡음이 없는 멜-스펙트로그램 간의 KL 다이버전스를 줄이는 방식으로 학습하고, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 텍스트를 컨디션으로 줄 때와 달리 같은 길이의 멜-스펙트로그램을 컨디션으로 주기 때문에 프레임 단위로 일대일 대응이 가능하고, 기존의 VITS 처럼 MAS를 사용하여 정렬할 필요가 없어 학습에 유리하도록 할 수 있다.

뿐만 아니라, 본 발명의 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 따르면, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 기존의 뉴럴넷 기반 음성 향상과 달리 잠음을 완전히 제거할 수 있도록 할 수 있다.

(52) CPC특허분류
G10L 25/18 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711153052
과제번호	2021-0-00456-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	원격 다자간 영상회의에서의 음성 품질 고도화 기술개발
기 여 율	1/1
과제수행기관명	한양대학교산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

음성 합성 구조 기반의 음성 향상 시스템(100)으로서,

잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 음성 인코더(speech encoder)(110); 및

상기 음성 인코더(110)로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더(120)를 포함하되,

상기 음성 인코더(110)는,

잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)을 입력받아 잠재 변수로 인코딩하는 제1 음성 인코더(111); 및

상기 제1 음성 인코더(111)의 잡음이 있는 음성의 멜-스펙트로그램에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 잠재 변수(latent variable)로 인코딩 하는 제2 음성 인코더(112)를 포함하여 구성하고,

상기 제1 음성 인코더(111)와 제2 음성 인코더(112)는,

잡음이 있는 음성의 멜-스펙트로그램과, 잡음이 없는 음성의 멜-스펙트로그램을 인코딩 처리하되, 동일한 구조의 인코더로 구성되며,

상기 음성 인코더(110)는,

잡음이 있는 음성의 멜-스펙트로그램과 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수(latent variable)로 인코딩 하되, 각각의 잠재 변수(latent variable)는 평균(mean)과 분산(variance)를 가진 정규분포(normal distribution)의 형태로 표현되고,

상기 음성 인코더(110)는,

잡음이 없는 음성에서 나온 잠재 변수(latent variable)을 정규화 흐름(normalizing flow)을 통해 변형하여 분포(distribution)를 더 복잡한 형태를 표현하고, 이를 다시 잡음이 없는 음성의 멜-스펙트로그램에서 인코딩된 잠재 변수와의 KL 다이버전스(divergence)를 줄이도록 기능하며,

상기 디코더(120)는,

상기 음성 인코더(110)를 통해 인코딩된 잠재 변수들이 KL 다이버전스(divergence)를 줄이도록 학습된 상태에서, 학습된 네트워크를 통해 생성 시에 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코더(120)를 통과하고, 이에 해당하는 잡음이 없는 음성이 생성되도록 기능하는 것을 특징으로 하는, 음성 합성 구조 기반의 음성 향상 시스템.

청구항 2

삭제

청구항 3

삭제

청구항 4

음성 합성 구조 기반의 음성 향상 방법으로서,

(1) 음성 인코더(110)가 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 단계;

(2) 디코더(120)가 상기 단계 (1)의 음성 인코더(110)로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 단계를 포함하되,

상기 단계 (1)에서의 음성 인코더(110)는,

잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)을 입력받아 잠재 변수로 인코딩하는 제1 음성 인코더(111); 및

상기 제1 음성 인코더(111)의 잡음이 있는 음성의 멜-스펙트로그램에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 잠재 변수(latent variable)로 인코딩 하는 제2 음성 인코더(112)를 포함하여 구성하고,

상기 제1 음성 인코더(111)와 제2 음성 인코더(112)는,

잡음이 있는 음성의 멜-스펙트로그램과, 잡음이 없는 음성의 멜-스펙트로그램을 인코딩 처리하되, 동일한 구조의 인코더로 구성되며,

상기 음성 인코더(110)는,

잡음이 있는 음성의 멜-스펙트로그램과 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수(latent variable)로 인코딩 하되, 각각의 잠재 변수(latent variable)는 평균(mean)과 분산(variance)를 가진 정규분포(normal distribution)의 형태로 표현되고,

상기 음성 인코더(110)는,

잡음이 없는 음성에서 나온 잠재 변수(latent variable)을 정규화 흐름(normalizing flow)을 통해 변형하여 분포(distribution)를 더 복잡한 형태를 표현하고, 이를 다시 잡음이 없는 음성의 멜-스펙트로그램에서 인코딩된 잠재 변수와의 KL 다이버전스(divergence)를 줄이도록 기능하며,

상기 디코더(120)는,

상기 음성 인코더(110)를 통해 인코딩된 잠재 변수들이 KL 다이버전스(divergence)를 줄이도록 학습된 상태에서, 학습된 네트워크를 통해 생성 시에 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코더(120)를 통과하고, 이에 해당하는 잡음이 없는 음성이 생성되도록 기능하는 것을 특징으로 하는, 음성 합성 구조 기반의 음성 향상 방법.

청구항 5

삭제

청구항 6

삭제

청구항 7

제4항에 따른 방법이 프로그램 명령어의 형태로 구현된 컴퓨터 판독 가능 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 관한 것으로서, 보다 구

체적으로는 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 하는 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 일실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0004] 음성 향상 기술은 그 자체로 모든 음성 통신 상황에서 사용할 수 있으며 음성 인식의 전처리 과정에서도 활용될 수 있어 사실상 음성 관련 모든 애플리케이션에서 적용할 수 있다. 특히 가전제품에도 음성 AI가 사용되는 흐름이 이어지면서 기존의 통신사, 인터넷 기업은 물론이고 스마트 TV, AI 스피커, 자동차 등의 하드웨어 사업에서도 적극적으로 음성 서비스를 사용하고 있다. 음성 인식 시장의 경우 2018년 750억 달러에서 2024년 2,150억 달러 규모로 성장할 것으로 전망된다. 이때, 음성 인식의 전처리로 사용되는 음성 향상의 경우에는 잡음이 없는 음성에 대해 학습이 요구된다.

[0006] 이러한 음성 향상 기술은 잡음이 섞인 음성에서 깨끗한 음성을 추정하는 기술로, 음성통신 분야에서는 음성의 명료도 향상에 도움을 주고, 음성 인식 에서는 전처리 기술로 이용하는 등 다양한 음성 관련 어플리케이션에 활용될 수 있는 중요한 기술이다. 초기 연구에서는 비음성 구간(노이즈만 있는 구간)에서 노이즈를 추정하여 그 정보를 바탕으로 노이즈를 제거하는 통계적 방법이 많이 사용되었으나, 이러한 기술은 노이즈가 시간에 따라 변하거나(non-stationary) 심하게 섞인 환경(low signal to ratio(SNR))에서는 성능이 저하되는 경향이 있었다.

[0008] 도 1은 기존의 종단형 음성 합성 모델(VITS)을 설명하기 위해 도시한 도면이다. 도 1에 도시된 바와 같이, 기존의 종단형 음성 합성 모델(VITS)는 음성 합성을 목적으로 하는 생성 모델로서, 텍스트를 컨디션으로 주고 이에 해당하는 음성의 파형을 생성해내고 있다. 이러한 기존의 뉴럴넷 기반 음성 향상에서는 잡음을 완전히 제거할 수 없는 문제가 있었다.

[0010] 전술한 배경 기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지 기술이라 할 수는 없다. 대한민국 공개특허 공보 제10-2017-0106312호가 선행기술 문헌으로 개시되고 있다.

발명의 내용

해결하려는 과제

[0011] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 음성 인코더와, 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더를 포함하여 구성함으로써, 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 하는, 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체를 제공하는 것을 그 목적으로 한다.

[0013] 또한, 본 발명은, 잡음이 있는 멜-스펙트로그램과 잡음이 없는 멜-스펙트로그램 간의 KL 다이버전스를 줄이는 방식으로 학습하고, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 텍스트를 컨디션으로 줄 때와 달리 같은 길이의 멜-스펙트로그램을 컨디션으로 주기 때문에 프레임 단위로 일대일 대응이 가능하고, 기존의 VITS 처럼 MAS를 사용하여 정렬할 필요가 없어 학습에 유리하도록 하는, 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체를 제공하는 것을 또 다른 목적으로 한다.

[0015] 뿐만 아니라, 본 발명은, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 기존의 뉴럴넷 기반 음성 향상과 달리 잡음을 완전히 제거할 수 있도록 하는, 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체를 제공하는 것을 또 다른 목적으로 한다.

[0017] 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술

적 과제들이 존재할 수 있다.

과제의 해결 수단

[0018] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 음성 합성 구조 기반의 음성 향상 시스템은,

[0019] 음성 합성 구조 기반의 음성 향상 시스템으로서,

[0020] 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 음성 인코더(speech encoder); 및

[0021] 상기 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더를 포함하는 것을 그 구성상의 특징으로 한다.

[0023] 바람직하게는, 상기 음성 인코더는,

[0024] 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)을 입력받아 잠재 변수로 인코딩하는 제1 음성 인코더; 및

[0025] 상기 제1 음성 인코더의 잡음이 있는 음성의 멜-스펙트로그램에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 잠재 변수(latent variable)로 인코딩 하는 제2 음성 인코더를 포함하여 구성할 수 있다.

[0027] 더욱 바람직하게는, 상기 음성 인코더는,

[0028] 잡음이 있는 음성의 멜-스펙트로그램과 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수(latent variable)로 인코딩 하되, 각각의 잠재 변수(latent variable)는 평균(mean)과 분산(variance)를 가진 정규분포(normal distribution)의 형태로 표현될 수 있다.

[0030] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 음성 합성 구조 기반의 음성 향상 방법은,

[0031] 음성 합성 구조 기반의 음성 향상 방법으로서,

[0032] (1) 음성 인코더가 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 단계;

[0033] (2) 디코더가 상기 단계 (1)의 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 단계를 포함하는 것을 그 구성상의 특징으로 한다.

[0035] 바람직하게는, 상기 단계 (1)에서의 음성 인코더는,

[0036] 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)을 입력받아 잠재 변수로 인코딩하는 제1 음성 인코더; 및

[0037] 상기 제1 음성 인코더의 잡음이 있는 음성의 멜-스펙트로그램에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 잠재 변수(latent variable)로 인코딩 하는 제2 음성 인코더를 포함하여 구성할 수 있다.

[0039] 더욱 바람직하게는, 상기 음성 인코더는,

[0040] 잡음이 있는 음성의 멜-스펙트로그램과 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수(latent variable)로 인코딩 하되, 각각의 잠재 변수(latent variable)는 평균(mean)과 분산(variance)를 가진 정규분포(normal distribution)의 형태로 표현될 수 있다.

[0042] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 음성 합성 구조 기반의 음성 향상 방법이 구현된 컴퓨터 판독 가능 매체는,

[0043] 본 발명의 특징에 따른 음성 합성 구조 기반의 음성 향상 방법이 프로그램 명령어의 형태로 구현된 것을 그 구

성상의 특징으로 한다.

발명의 효과

[0044] 본 발명에서 제안하고 있는 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 따르면, 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 음성 인코더와, 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더를 포함하여 구성함으로써, 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 할 수 있다.

[0046] 또한, 본 발명의 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 따르면, 잡음이 있는 멜-스펙트로그램과 잡음이 없는 멜-스펙트로그램 간의 KL 다이버전스를 줄이는 방식으로 학습하고, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 텍스트를 컨디션으로 줄 때와 달리 같은 길이의 멜-스펙트로그램을 컨디션으로 주기 때문에 프레임 단위로 일대일 대응이 가능하고, 기존의 VITS 처럼 MAS를 사용하여 정렬할 필요가 없어 학습에 유리하도록 할 수 있다.

[0048] 뿐만 아니라, 본 발명의 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체에 따르면, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 기존의 뉴럴넷 기반 음성 향상과 달리 잡음을 완전히 제거할 수 있도록 할 수 있다.

도면의 간단한 설명

[0049] 도 1은 기존의 종단형 음성 합성 모델(VITS)을 설명하기 위해 도시한 도면.

도 2는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 구성을 기능블록으로 도시한 도면.

도 3은 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 음성 인코더의 구성을 기능블록으로 도시한 도면.

도 4는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 모델 구조를 개략적으로 도시한 도면.

도 5는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 방법의 흐름을 도시한 도면.

발명을 실시하기 위한 구체적인 내용

[0050] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0052] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0054] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.

[0056] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.

[0058] 도 2는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 구성을 기능블록으로 도시한 도면이다. 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템(100)은, 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 음성 인코더(speech encoder)(110)와, 음성 인코더(110)로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더(120)를 포함하여 구성될 수 있다. 이하에서는 첨부된 도면을 참조하여 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 구체적인 구성에 대해 상세히 설명하기로 한다.

[0060] 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템을 설명하기에 앞서 음성 합성 모델(VITS)에 대해 설명하기로 한다.

[0062] 먼저, VITS(Variational Inference with adversarial learning for end-to-end Text-to-Speech)는 end-to-end 음성 합성 모델로, 파형(waveform)을 직접 생성해내는 점과 양질의 샘플을 만들어내는 것을 특징으로 하여 음성 합성 분야에서는 다년간 주요하게 사용되고 있다. VITS는 조건적 VAE(Variational AutoEncoder)를 기반으로 하고 있으며, VAE는 생성 모델의 하나로 학습한 데이터와 유사한 데이터를 생성해 내는 것을 목표로 한다. 학습 과정에서는 인코더와 디코더 구조로 되어 데이터를 인코더를 통해 잠재 변수의 형태로 인코딩하고 이를 다시 디코딩하여 기존의 데이터로 복구하는 방식으로 학습한다. 이는 즉 ELBO(Evidence Lower Bound)를 최적화 하는 과정으로 볼 수 있는데 ELBO는 아래의 [수학식 1]과 같이 나타낼 수 있다.

수학식 1

[0064] $$E_{q_\phi(z|x)}[\log(p(x|z))-KL(q_\phi(z|x)||p(z))+KL(q_\phi(z|x)||p(z|x))$$

[0066] 이와 같이, ELBO에서 첫 번째 기간(term)을 복원 기간(Reconstruction term)이라 하고 이를 최대화하여 입력 데이터가 모델을 통과했을 때 복원이 되도록 한다. 두 번째 기간(term)은 정규화 기간(Regularization term)으로 이전(prior) p(z)와 이상적인 샘플링 함수 간의 KL loss로 이전(prior)과 유사한 값을 샘플링 하도록 한다. 마지막은 이상적 샘플링 함수와 실제 샘플링 함수 간의 거리로 측정할 수 없기 때문에 앞선 두 기간(term)을 최적화 하는 것이 VAE의 목표이다.

[0068] 이러한 VITS의 경우 조건적 VAE를 기반으로 하여 텍스트(Text)를 컨디션으로 주고 학습하여 생성 시에 텍스트(text)를 넣으면 음성이 합성되도록 한다. 그 과정에서 위와 같이 reconstruction loss, KL loss를 사용하며 추가로 표현력을 높이기 위한 정규화 흐름(normalizing flow) 구조와 텍스트와 멜-스펙트트로그램 간의 정렬(align)을 해결하기 위한 MAS(Monotonic Alignment Search)를 사용하여 학습한다.

[0070] 기존의 VITS는 음성 합성을 목적으로 하는 생성 모델로 텍스트(text)를 컨디션으로 주고 이에 해당하는 파형(waveform)을 생성해내는 것을 목적으로 한다. 그러나 텍스트 대신 잡음이 섞인 파형의 멜-스펙트로그램을 컨디션으로 주어 학습하는 본 발명의 경우 음성 향상을 목적으로 하고 있으며, VITS의 장점인 고품질의 생성 결과와 end-to-end로 결과를 만들어 낼 수 있는 특징을 그대로 활용할 수 있다. 또한 생성모델을 기반으로 하여 잡음이 없는 파형(waveform)을 생성하도록 학습되므로 기존의 음성 향상 모델들과 달리 잡음이 완전히 제거된 결과를 얻을 수 있다.

[0072] 도 3은 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 음성 인코더의 구성을 기능블록으로 도시한 도면이고, 도 4는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 모델 구조를 개략적으로 도시한 도면이다.

[0074] 음성 인코더(speech encoder)(110)는, 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 하는 구성이다. 이러한 음성 인코더(110)는 도 3 및 도 4에 각각 도시된 바와 같이, 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)을 입력받아 잠재 변수로 인코딩하는 제1 음성 인코더(111)와, 제1 음성 인코더(111)의 잡음이 있는 음성의 멜-스펙트로그램에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 잠재 변수(latent variable)로 인코딩 하는 제2 음성 인코더(112)를 포함하여 구성할 수 있다. 여기서, 제1 음성 인코더(111)와 제2 음성 인코더(112)는 동일한

구조의 인코더로서, 잡음이 있는 음성의 멜-스펙트로그램과, 잡음이 있는 음성의 멜-스펙트로그램을 인코딩 처리하기 위한 구성이다.

[0076] 음성 인코더(110)는 잡음이 있는 음성의 멜-스펙트로그램과 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수(latent variable)로 인코딩 하되, 각각의 잠재 변수(latent variable)는 평균(mean)과 분산(variance)를 가진 정규분포(normal distribution)의 형태로 표현될 수 있다. 이때, 잡음이 없는 음성에서 나온 잠재 변수(latent variable)을 정규화 흐름(normalizing flow)을 통해 변형하여 분포(distribution)를 더 복잡한 형태를 표현할 수 있도록 한다. 그리고 이를 다시 잡음이 없는 음성의 멜-스펙트로그램에서 인코딩된 잠재 변수와의 KL divergence를 줄이도록 한다.

[0078] 디코더(120)는, 음성 인코더(110)로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 구성이다. 이러한 디코더(120)는 음성 인코더(110)를 통해 인코딩된 잠재 변수들이 KL divergence를 줄이도록 학습된 상태에서, 학습된 네트워크를 통해 생성 시에 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코더(120)를 통과하고, 이에 해당하는 잡음이 없는 음성이 생성되도록 한다.

[0080] 도 4는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 모델 구조를 개략적으로 나타내고 있다. 이하에서는 첨부된 도면을 참조하여 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템의 구체적인 실시예를 설명하기로 한다.

[0082] 먼저, 모델 구조로, 음성 합성 구조 기반의 음성 향상 시스템(100)은 음성 인코더(110)와 디코더(120)로 구성될 수 있다. 음성 인코더(110)는 확장된 회선(Dilated convolution)과 스킵 컨넥션(skip connection)으로 구성된 non-causual WaveNet residual block을 통해 멜-스펙트로그램을 인코딩 한다. 결과로 나온 잠재 변수(latent variable)는 평균(mean)과 분산(variance)을 통한 정규분포(normal distribution)로 표현된다. 잡음이 없는 멜-스펙트로그램(Clean speech mel-spectrogram)과 잡음이 있는 멜-스펙트로그램(Noisy speech mel-spectrogram) 모두 같은 구조의 음성 인코더(111, 112)를 통해 latent로 맵핑(mapping) 된다.

[0084] 또한, 디코더(120)는 컨볼루션 레이어(Convolution layer)에 이어진 Multi-receptive field fusion module(MRF) 로 이루어진 Hifi-GAN V1 generator 기반으로 입력(input) z를 파형(waveform)으로 변환해준다. 또한, 도시되지는 안았지만, 판별기(Discriminator)는 여러 개의 Markovian window-based discriminator들로 구성된 Multi-period discriminator를 이용하여 다양한 period에 대해서 GAN 방식으로 training 하는데 사용한다.

[0086] 학습 방법에는, 두 가지 Main Loss가 사용된다. 이 모델은 두 가지 loss term을 이용하여 학습하는데, 첫 번째는 reconstruction loss로 디코더(120)를 통해 얻은 파형(waveform)을 멜-스펙트로그램으로 변환한 뒤에 입력(input)으로 넣은 잡음이 없는 음성과 비교하여 복원 정도를 L1 loss로 표현한다. 두 번째는 KL loss로 기존 VAE와 다르게 컨디션으로 준 잡음이 있는 음성과 잡음이 없는 음성 간의 KL divergence를 Loss로 사용한다. 이때 가우시안으로 가정한 prior를 정규화 흐름을 활용하여 distribution을 complex하게 바꾸어 표현력을 늘리도록 하였다. 이를 통해 음성 생성 시에 잡음이 있는 음성을 적절한 잠재 변수의 형태로 인코딩하고 그 결과가 잡음이 없는 음성의 데이터 분포와 매우 유사해지도록 하여 다시 디코더(120)를 통해 음성으로 변환하면서 잡음이 없는 음성을 생성하도록 한다. 이를 아래의 [수학식 2]로 표현하면 다음과 같다.

수학식 2

$$L_{recon} = ||x - \hat{x}_{mel}||_1$$

$$L_{kl} = \log q_\phi(z|x) - \log p_\theta(z|c)$$

$$z \sim \log q_\phi(z|x) = N(z; \mu_\theta(x), \sigma_\theta(x))$$

$$p_\theta(z|c) = N(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|$$

[0088]

[0090] 또한, 적대적 훈련으로서, GAN 방식의 적대적 훈련은 판별기(discriminator) D와 디코더(decoder) G를 이용하여

훈련하는 방식인데, 다음과 같은 loss를 주어 G에서 discriminator가 진짜와 구분하지 못하는 sample을 만들도록 한다. 이를 [수학식 3]으로 표현하면 다음과 같다.

수학식 3

[0092]

$$L_{adv}(D)=E_{y,z}[D(y)-1)^2+(D(G(z)))^2]$$

$$L_{adv}(G)=E_z[D(G(z)-1)^2]$$

[0094] 검증결과를 보면, 아래의 [표 1]에 도시된 바와 같이, 새로운 기법은 기존의 기술(SEGAN) 에 비해 높은 PESQ를 얻을 수 있음을 확인할 수 있다. 실험은 다화자 음성 향상 데이터 셋인 Valentini Dataset을 사용하여 진행되었으며, 모델의 input으로 멜-스펙트로그램을 사용하여 phase 정보를 얻을 수 없었기에, 결과로 나온 음성에 phase 값을 잡음이 있는 음성의 phase로 교체하여 PESQ score를 계산하였다.

표 1

[0096]

Model	PESQ
Noisy	1.97
SEGAN	2.16
Proposed	2.26

[0098] 도 5는 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 방법의 흐름을 도시한 도면이다. 도 5에 도시된 바와 같이, 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 방법은, 음성 인코더가 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 단계(S110), 및 디코더가 단계 S110의 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 단계(S120)를 포함하여 구현될 수 있다.

[0100] 단계 S110에서는, 음성 인코더(110)가 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 각각의 잠재 변수(latent variable)로 인코딩 한다. 이러한 단계 S110에서의 음성 인코더(110)는 도 3 및 도 4에 각각 도시된 바와 같이, 잡음이 있는 음성의 멜-스펙트로그램(Noisy Speech Mel-Spectrogram)을 입력받아 잠재 변수로 인코딩하는 제1 음성 인코더(111)와, 제1 음성 인코더(111)의 잡음이 있는 음성의 멜-스펙트로그램에 해당하는 잡음이 없는 음성의 멜-스펙트로그램(Clean Speech Mel-Spectrogram)을 잠재 변수(latent variable)로 인코딩 하는 제2 음성 인코더(112)를 포함하여 구성할 수 있다. 여기서, 제1 음성 인코더(111)와 제2 음성 인코더(112)는 동일한 구조의 인코더로서, 잡음이 있는 음성의 멜-스펙트로그램과, 잡음이 있는 음성의 멜-스펙트로그램을 인코딩 처리하기 위한 구성이다.

[0102] 또한, 음성 인코더(110)는 잡음이 있는 음성의 멜-스펙트로그램과 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수(latent variable)로 인코딩 하되, 각각의 잠재 변수(latent variable)는 평균(mean)과 분산(variance)를 가진 정규분포(normal distribution)의 형태로 표현될 수 있다. 이때, 잡음이 없는 음성에서 나온 잠재 변수(latent variable)을 정규화 흐름(normalizing flow)을 통해 변형하여 분포(distribution)를 더 복잡한 형태를 표현할 수 있도록 한다. 그리고 이를 다시 잡음이 없는 음성의 멜-스펙트로그램에서 인코딩된 잠재 변수와의 KL divergence를 줄이도록 한다.

[0104] 단계 S120에서는, 디코더(120)가 단계 S110의 음성 인코더(110)로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성한다. 이러한 단계 S120에서의 디코더(120)는 음성 인코더(110)를 통해 인코딩된 잠재 변수들이 KL divergence를 줄이도록 학습된 상태에서, 학습된 네트워크를 통해 생성 시에 잡음이 섞인 음성에 대한 멜

-스펙트로그램을 컨디션으로 하여 디코더(120)를 통과하고, 이에 해당하는 잡음이 없는 음성이 생성되도록 한다.

[0106] 한편, 본 발명은 다양한 통신 단말기로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터에서 판독 가능한 매체를 포함할 수 있다. 예를 들어, 컴퓨터에서 판독 가능한 매체는, 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 장치를 포함할 수 있다.

[0108] 이와 같은 컴퓨터에서 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이때, 컴퓨터에서 판독 가능한 매체에 기록되는 프로그램 명령은 본 발명을 구현하기 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 예를 들어, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0110] 이와 같은 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체는, 음성 합성 모델인 VITS(Variational Inference with adversarial learning for end-to-end Text-to-Speech) 구조를 음성 향상에 활용하여 생성모델 기반의 음성 향상 기법을 제시하고, 이 생성모델을 기반으로 한 기법은 기존의 뉴럴넷 기반 음성향상과 달리 잡음을 완전히 제거할 수 있으면서도 성능을 끌어올리는 효과를 얻을 수 있으며, 조건적 VAE 기반의 구조로서 컨디션(condition)에 따른 잠재 변수(latent variable)의 인코딩과 디코딩을 학습하여 컨디션에 따른 결과를 생성할 수 있도록 한다.

[0112] 또한, 본 발명의 음성 향상 시스템은 음성 향상 기술이 적용될 수 있는 대부분의 음성 관련 기술에 모두 적용될 수 있는 기술로서, 잡음이 심한 음성의 잡음을 완전히 제거하여 품질이 좋은 음성을 생성해낼 수 있어 음성 인식에도 유용하며, 거의 모든 통신 상황에서 적용할 수 있어 관련 상품과 서비스에 사용할 수 있다. 또한, 소프트웨어 프로그램에 관한 것으로 프로그램 배포를 통한 대량 생산에 용이하며, 기존의 상품 및 서비스에 적용되고 있는 음성 관련 기술에 추가로 적용할 수 있어 산업 적용에 용이하다.

[0114] 또한, 관련된 서비스를 제공하는 대표적인 회사는 네이버, 카카오, SKT, LG 등이 있으며 좋은 품질의 음성 향상 모델을 필요로 하는 기업의 기술이전 가능성도 높을 것으로 예상되고, 음성 인식을 위한 전처리로 사용될 수 있어, AI 스피커, 스마트폰 등 음성 인식을 요하는 대부분의 디바이스에 적용이 가능하며 화상회의, 개인방송, 안내방송 등 잡음이 방해될 여지가 있는 다른 산업에도 적용 가능하다.

[0116] 상술한 바와 같이, 본 발명의 일실시예에 따른 음성 합성 구조 기반의 음성 향상 시스템, 방법 및 컴퓨터 판독 가능 매체는, 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 음성 인코더와, 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 디코더를 포함하여 구성함으로써, 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 할 수 있으며, 또한, 잡음이 있는 멜-스펙트로그램과 잡음이 없는 멜-스펙트로그램 간의 KL 다이버전스를 줄이는 방식으로 학습하고, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 텍스트를 컨디션으로 줄 때와 달리 같은 길이의 멜-스펙트로그램을 컨디션으로 주기 때문에 프레임 단위로 일대일 대응이 가능하고, 기존의 VITS 처럼 MAS를 사용하여 정렬할 필요가 없어 학습에 유리하도록 할 수 있으며, 특히, 학습 이후 음성 합성 모델인 VITS 구조를 음성 향상에 활용하여 생성모델 기반으로 잡음이 섞인 음성에서 잡음이 없는 음성을 생성할 수 있도록 함으로써, 기존의 뉴럴넷 기반 음성 향상과 달리 잡음을 완전히 제거할 수 있도록 할 수 있게 된다.

[0118] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0120] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으

로 해석되어야 한다.

부호의 설명

[0121] 100: 본 발명의 일실시예에 따른 음성 향상 시스템

110: 음성 인코더

111: 제1 음성 인코더

112: 제2 음성 인코더

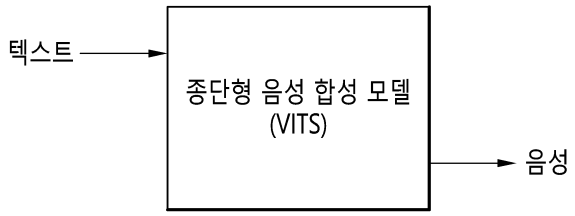
120: 디코더

S110: 음성 인코더가 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 단계

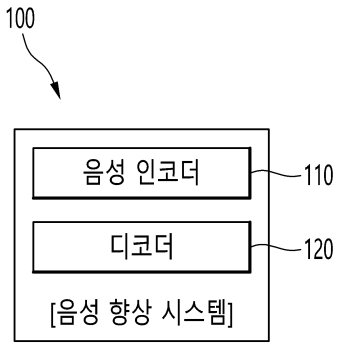
S120: 디코더가 단계 S110의 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 단계

도면

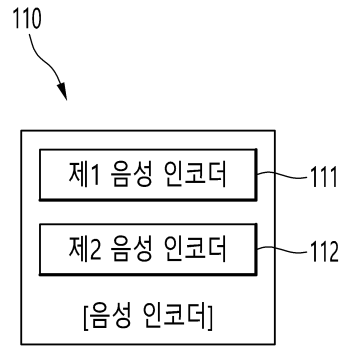
도면1



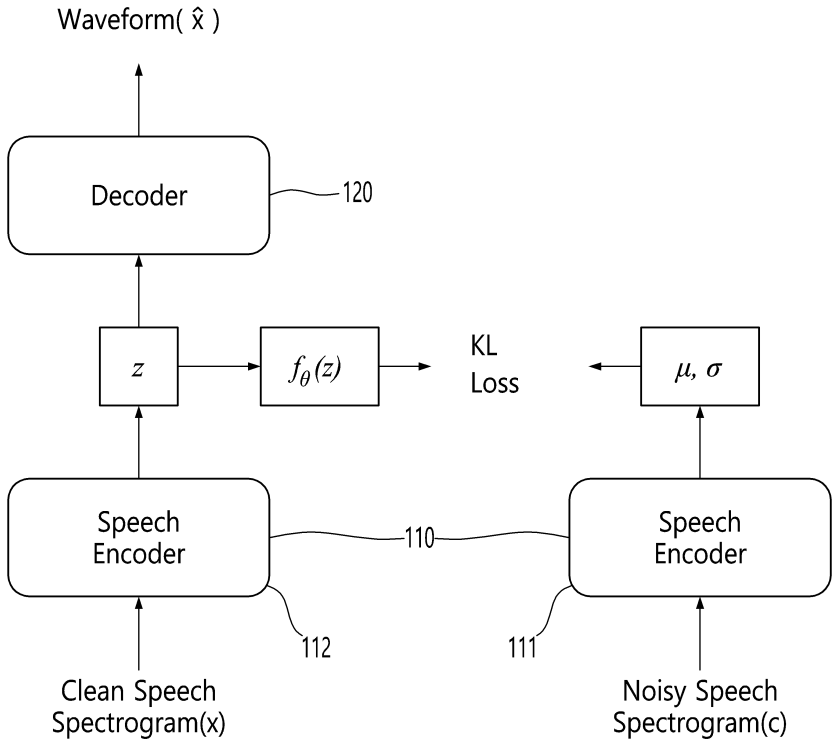
도면2



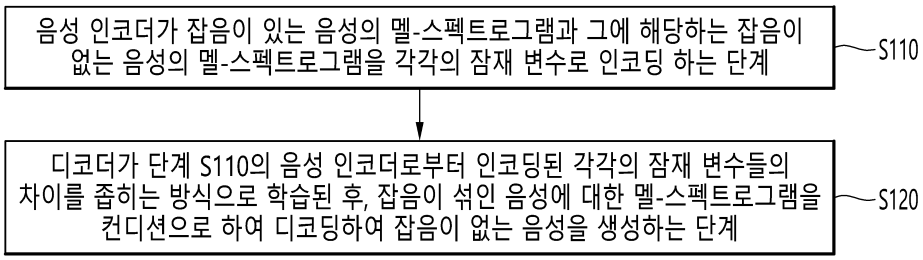
도면3


110
제1 음성 인코더
111
제2 음성 인코더
112
[음성 인코더]

도면4


Waveform(x̂)
Decoder
120
z
$f_\theta(z)$
KL
Loss
μ, σ
Speech
Encoder
110
Speech
Encoder
112
111
Clean Speech
Spectrogram(x)
Noisy Speech
Spectrogram(c)

도면5


음성 인코더가 잡음이 있는 음성의 멜-스펙트로그램과 그에 해당하는 잡음이 없는 음성의 멜-스펙트로그램을 각각의 잠재 변수로 인코딩 하는 단계
S110
디코더가 단계 S110의 음성 인코더로부터 인코딩된 각각의 잠재 변수들의 차이를 좁히는 방식으로 학습된 후, 잡음이 섞인 음성에 대한 멜-스펙트로그램을 컨디션으로 하여 디코딩하여 잡음이 없는 음성을 생성하는 단계
S120