(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0097055
(43) 공개일자 2025년06월30일

(51) 국제특허분류(Int. Cl.)
G10L 13/033 (2013.01) *G10L 17/02* (2013.01)
G10L 19/00 (2006.01) *G10L 21/003* (2013.01)
G10L 25/30 (2013.01)
(52) CPC특허분류
G10L 13/033 (2013.01)
G10L 17/02 (2013.01)
(21) 출원번호 10-2023-0187924
(22) 출원일자 2023년12월21일
심사청구일자 2023년12월21일

(71) 출원인
주식회사 앨리스앤원더랩스
울산광역시 울주군 언양읍 유니스트길 50, 251동 322호(울산과학기술원)
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
이지민
수원시 영통구 영흥숲길 50
조형주
경기도 성남시 분당구 내정로166번길 7-6
류동민
수원시 영통구 영흥숲길 50
(74) 대리인
특허법인엠에이피에스

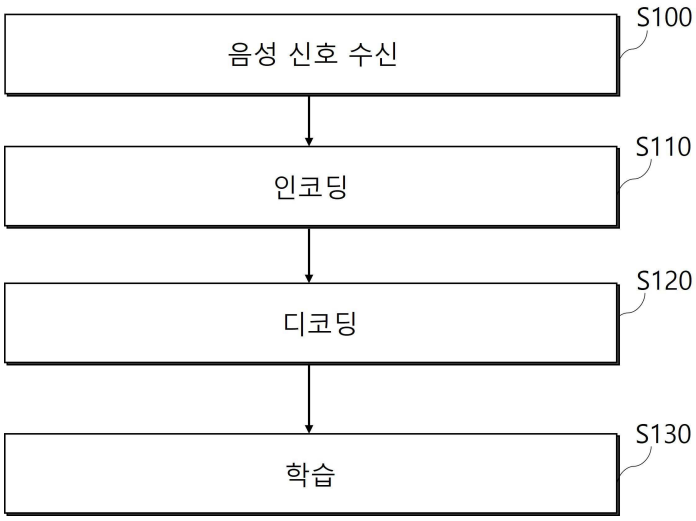
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **음성 변환 방법 및 장치**

(57) 요 약

본 발명의 실시예에 따른 음성 변환 방법은 소스 음성 신호와 타겟 음성 신호를 수신하는 음성 신호 수신 단계; 상기 소스 음성 신호와 상기 타겟 음성 신호를 인코더에 입력하고, 상기 인코더로부터 상기 소스 음성 신호와 상기 타겟 음성 신호 각각의 컨텐츠 벡터와 스타일 벡터가 출력되는 인코딩 단계; 출력된 상기 컨텐츠 벡터 및 스타일 벡터를 디코더에 입력하고, 상기 디코더로부터 합성 음성 신호가 출력되는 디코딩 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 19/0017 (2013.01)
G10L 21/003 (2013.01)
G10L 25/30 (2013.01)

명 세 서

청구범위

청구항 1

음성 변환 장치에 의한 음성 변환 방법에 있어서,

소스 음성 신호와 타겟 음성 신호를 수신하는 음성 신호 수신 단계;

상기 소스 음성 신호와 상기 타겟 음성 신호를 인코더에 입력하고, 상기 인코더로부터 상기 소스 음성 신호와 상기 타겟 음성 신호 각각의 컨텐츠 벡터와 스타일 벡터를 출력하는 인코딩 단계;

출력된 상기 컨텐츠 벡터 및 스타일 벡터를 디코더에 입력하고, 상기 디코더로부터 합성 음성 신호를 출력하는 디코딩 단계; 를 포함하는, 음성 변환 방법.

청구항 2

제1항에 있어서,

상기 인코더는

음성 신호가 입력되면 상기 음성 신호를 시계열에 따른 프레임 윈도우로 분리하고, 분리된 프레임 윈도우 별로 위치 값이 부여된 포지셔널 임베딩 정보가 포함되는 임베딩 벡터를 출력하는 음성 임베딩 레이어를 포함하는, 음성 변환 방법.

청구항 3

제2항에 있어서,

상기 인코더는

상기 임베딩 벡터로부터 공유 벡터를 출력하는 트랜스포머 인코더,

상기 공유 벡터로부터 상기 컨텐츠 벡터를 출력하는 컨텐츠 인코더 및

상기 공유 벡터로부터 상기 스타일 벡터를 출력하는 스타일 인코더를 더 포함하는, 음성 변환 방법.

청구항 4

제1항에 있어서,

상기 트랜스포머 인코더는 상기 임베딩 벡터의 적어도 일부를 마스킹하는 자기 지도 학습을 통해 학습되는 것인, 음성 변환 방법.

청구항 5

제1항에 있어서,

상기 합성 음성 신호는

상기 소스 음성 신호의 컨텐츠 벡터와 상기 타겟 음성 신호의 스타일 벡터가 합성된 음성 신호인 것인, 음성 변환 방법.

청구항 6

제1항에 있어서,

상기 디코더를 학습시키는 학습 단계; 를 더 포함하고,

상기 학습 단계는

학습 소스 음성 신호와 학습 타겟 음성 신호가 상기 인코더에 입력되는 입력 단계;

상기 인코더가 상기 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩 단계;

상기 인코더가 상기 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩 단계;

상기 디코더가 상기 소스 컨텐츠 벡터, 상기 소스 스타일 벡터, 상기 타겟 컨텐츠 벡터 및 상기 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩 단계; 및

상기 학습 합성 음성 신호에 손실 함수를 적용하고, 상기 손실 함수에 의한 손실이 적어지도록 상기 디코더를 튜닝하는 튜닝 단계; 를 포함하는, 음성 변환 방법.

청구항 7

제6항에 있어서,

상기 학습 소스 음성 신호와 상기 학습 타겟 음성 신호는 같은 발화자에 의해 생성된 음성 신호인, 음성 변환 방법.

청구항 8

제7항에 있어서,

상기 손실 함수는 상기 소스 스타일 벡터와 상기 타겟 스타일 벡터 간의 차이가 적어지도록 설정된 것인, 음성 변환 방법.

청구항 9

음성 변환 모델 학습 장치에 의한 음성 변환 모델 학습 방법에 있어서,

학습 소스 음성 신호와 학습 타겟 음성 신호가 인코더에 입력되는 학습 음성 신호 입력 단계;

상기 인코더가 상기 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩 단계;

상기 인코더가 상기 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩 단계;

디코더가 상기 소스 컨텐츠 벡터, 상기 소스 스타일 벡터, 상기 타겟 컨텐츠 벡터 및 상기 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩 단계; 및

상기 학습 합성 음성 신호에 손실 함수를 적용하여 상기 디코더를 튜닝하는 튜닝 단계; 를 포함하는 음성 변환 모델 학습 방법.

청구항 10

제9항에 있어서,

상기 학습 소스 음성 신호와 상기 학습 타겟 음성 신호는 같은 발화자에 의해 생성된 음성 신호인, 음성 변환 모델 학습 방법.

청구항 11

제9항에 있어서,

상기 손실 함수는

상기 학습 소스 음성 신호 또는 상기 학습 타겟 음성 신호와 상기 학습 합성 음성 신호 간의 차이를 계산하는 신호 복원 손실 함수를 포함하는 음성 변환 모델 학습 방법.

청구항 12

제9항에 있어서,

상기 학습 합성 음성 신호가 복원 인코더에 입력되고, 상기 복원 인코더가 상기 학습 합성 음성 신호로부터 복원 컨텐츠 벡터와 복원 스타일 벡터를 출력하는 복원 인코딩 단계; 를 더 포함하고,

상기 튜닝 단계는 상기 학습 합성 음성 신호, 상기 복원 컨텐츠 벡터 및 상기 복원 스타일 벡터 중 적어도 하나 이상에 상기 손실 함수를 적용하여 상기 디코더를 튜닝하는 단계인, 음성 변환 모델 학습 방법,

청구항 13

제12항에 있어서,

상기 손실 함수는

상기 소스 컨텐츠 벡터 또는 상기 타겟 컨텐츠 벡터와 상기 복원 컨텐츠 벡터 간의 차이를 계산하는 컨텐츠 복원 손실 함수를 포함하는 음성 변환 모델 학습 방법.

청구항 14

제12항에 있어서,

상기 손실 함수는

상기 소스 스타일 벡터 또는 상기 타겟 스타일 벡터와 상기 복원 스타일 벡터 간의 차이를 계산하는 스타일 복원 손실 함수를 포함하는 음성 변환 모델 학습 방법.

청구항 15

제9항에 있어서,

상기 디코더는 생성적 대립 신경망(Generative Adversarial Network)에 기반하는 것인, 음성 변환 모델 학습 방법.

청구항 16

제9항에 있어서,

상기 디코더는 디퓨전 모델(Diffusion model)에 기반하는 것인 음성 변환 모델 학습 방법.

청구항 17

음성 변환 장치에 있어서,

적어도 하나의 프로세서; 및

컴퓨터 프로그램 코드를 포함하는 적어도 하나의 메모리를 포함하고,

상기 적어도 하나의 메모리와 상기 컴퓨터 프로그램 코드는 상기 적어도 하나의 프로세서를 통해 상기 장치가,

소스 음성 신호와 타겟 음성 신호를 수신하는 음성 신호 수신부;

상기 소스 음성 신호와 상기 타겟 음성 신호를 인코더에 입력하고, 상기 인코더로부터 상기 소스 음성 신호와 상기 타겟 음성 신호 각각의 컨텐츠 벡터와 스타일 벡터를 출력하는 인코딩부;

출력된 상기 컨텐츠 벡터 및 스타일 벡터를 디코더에 입력하고, 상기 디코더로부터 합성 음성 신호를 출력하는 디코딩부; 를 포함하는, 음성 변환 장치.

청구항 18

음성 변환 모델 학습 장치에 있어서,

적어도 하나의 프로세서; 및

컴퓨터 프로그램 코드를 포함하는 적어도 하나의 메모리를 포함하고,

상기 적어도 하나의 메모리와 상기 컴퓨터 프로그램 코드는 상기 적어도 하나의 프로세서를 통해 상기 장치가,

학습 소스 음성 신호와 학습 타겟 음성 신호가 인코더에 입력되는 학습 음성 신호 입력부;

상기 인코더가 상기 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩부;

상기 인코더가 상기 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩부;

디코더가 상기 소스 컨텐츠 벡터, 상기 소스 스타일 벡터, 상기 타겟 컨텐츠 벡터 및 상기 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩부; 및

상기 학습 합성 음성 신호에 손실 함수를 적용하여 상기 디코더를 튜닝하는 튜닝부; 를 포함하는 음성 변환 모델 학습 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 입력되는 음성 신호를 변환하는 변환 모델에 기초하여 음성 신호를 변환하는 방법 및 장치와 음성 변환 모델을 학습시키는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 몇 년간, 음성 처리 기술은 많은 발전을 이루었다. 초기의 단순한 음성 인식 시스템에서부터 시작하여, 오늘날에는 자연스러운 음성 합성, 음성 변환, 그리고 언어 이해에 이르기까지, 다양한 분야에서 혁신이 이루어지고 있다. 이러한 발전은 딥 러닝과 같은 고급 머신 러닝 기법들의 발전과 뗄 수 없는 관계가 있으며, 대규모 데이터셋의 사용과 더불어 더욱 정교해진 알고리즘의 개발이 주요 원인으로 작용하였다.

[0003] 음성 변환 기술은 다양한 응용 분야에서 중요한 역할을 한다. 예를 들어, 영화 및 게임 산업에서는 캐릭터에 생동감을 부여하기 위해 다양한 목소리를 생성하거나 변환하는 데 이 기술이 사용된다. 또한, 개인화된 음성 비서, 음성 기반 번역 서비스, 가상 현실 등에서도 맞춤형 음성 경험을 제공하기 위한 핵심 기술로 자리잡고 있다.

[0004] 음성 스타일 변환은 단순히 텍스트를 음성으로 변환하는 것을 넘어, 특정 인물의 음성 스타일을 모방하거나 변경하는 고급 기술이다. 이는 발화자의 특성을 포착하고, 그 스타일을 다른 음성 콘텐츠에 적용함으로써, 개성 있는 음성 출력을 가능하게 한다. 이러한 변환 과정은 억양, 강세, 음색 등과 같은 다양한 음성 특성을 고려하여 이루어진다.

[0005] 음성 변환 기술은 변환된 음성의 자연스러움과 정확성을 동시에 달성하는 것이 중요하다. 하지만, 현재 기술은 이러한 목표에 대해 아직 완벽하지 않으며, 특히 다양한 언어와 방언에서의 정확한 스타일 변환은 여전히 개선의 여지가 크다. 또한, 실시간 변환과 저비용 고효율 처리 등의 문제도 해결해야 할 과제로 남아있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 사용자가 입력한 음성 신호의 컨텐츠를 유지한 채, 스타일을 변형하는 음성 변환 모델을 제공하고자 한다.

[0007] 또한, 본 발명은 음성 변환 모델을 통해 생성된 합성 음성 신호가 변환 대상이 되는 음성의 스타일을 자연스럽고 높은 정확도로 적용할 수 있는 음성 변환 방법을 제공하고자 한다.

[0008] 또한, 본 발명은 자연스러운 음성 변환을 가능하도록 하는 음성 변환 모델을 학습시키는 방법을 제공하고자 한다.

[0009] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0010] 상술한 기술적 과제를 달성하기 위한 수단으로서 본 발명의 일 실시예는, 음성 변환 장치에 의한 음성 변환 방법에 있어서, 소스 음성 신호와 타겟 음성 신호를 수신하는 음성 신호 수신 단계; 상기 소스 음성 신호와 상기 타겟 음성 신호를 인코더에 입력하고, 상기 인코더로부터 상기 소스 음성 신호와 상기 타겟 음성 신호 각각의 컨텐츠 벡터와 스타일 벡터가 출력되는 인코딩 단계; 출력된 상기 컨텐츠 벡터 및 스타일 벡터를 디코더에 입력하고, 상기 디코더로부터 합성 음성 신호가 출력되는 디코딩 단계; 를 포함할 수 있다.

[0011] 또한, 본 발명의 일 실시예는, 음성 변환 모델 학습 장치에 의한 음성 변환 모델 학습 방법에 있어서, 학습 소스 음성 신호와 학습 타겟 음성 신호가 인코더에 입력되는 학습 음성 신호 입력 단계; 상기 인코더가 상기 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩 단계; 상기 인코더가 상기 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩 단계; 디코더가 상기 소스 컨텐츠 벡터, 상기 소스 스타일 벡터, 상기 타겟 컨텐츠 벡터 및 상기 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩 단계; 및 상기 학습 합성 음성 신호에 손실 함수를 적용하여 상기 디코더를 튜닝하는 튜닝 단계; 를 포함할 수 있다.

[0012] 상술한 과제 해결 수단은 단지 예시적인 것으로서, 본 발명을 제한하려는 의도로 해석되지 않아야 한다. 상술한 예시적인 실시예 외에도, 도면 및 발명의 상세한 설명에 기재된 추가적인 실시예가 존재할 수 있다.

발명의 효과

[0013] 전술한 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 본 발명은 사용자가 입력한 음성 신호의 컨텐츠를 유지한 채, 스타일을 변형하는 음성 변환 모델을 제공할 수 있다.

[0014] 또한, 본 발명은 음성 변환 모델을 통해 생성된 합성 음성 신호가 변환 대상이 되는 음성의 스타일을 자연스럽고 높은 정확도로 적용할 수 있는 음성 변환 방법을 제공할 수 있다.

[0015] 또한, 본 발명은 자연스러운 음성 변환을 가능하도록 하는 음성 변환 모델을 학습시키는 방법을 제공할 수 있다.

도면의 간단한 설명

[0016] 도 1내지 도 3은 본 발명의 일 실시예에 따른 음성 변환 방법을 설명하기 위한 흐름도이다.

도 4와 도 5는 본 발명의 일 실시예에 따른 음성 변환 모델 학습 방법을 설명하기 위한 흐름도이다.

도 6과 도 7은 본 발명의 일 실시예에 따른 음성 변환 모델 학습 방법을 설명하기 위한 흐름도이다.

도 8은 본 발명의 일 실시예에 따른 음성 변환 장치를 설명하기 위한 구성도이다.

도 9는 본 발명의 일 실시예에 따른 음성 변환 모델 학습 장치를 설명하기 위한 구성도이다.

발명을 실시하기 위한 구체적인 내용

[0017] 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.

[0018] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 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.

[0019] 본 명세서에 있어서 '부(部)'란, 하드웨어에 의해 실현되는 유닛(unit), 소프트웨어에 의해 실현되는 유닛, 양방을 이용하여 실현되는 유닛을 포함한다. 또한, 1 개의 유닛이 2 개 이상의 하드웨어를 이용하여 실현되어도 되고, 2 개 이상의 유닛이 1 개의 하드웨어에 의해 실현되어도 된다.

[0020] 본 명세서에 있어서 단말 또는 디바이스가 수행하는 것으로 기술된 동작이나 기능 중 일부는 해당 단말 또는 디

바이스와 연결된 서버에서 대신 수행될 수도 있다. 이와 마찬가지로, 서버가 수행하는 것으로 기술된 동작이나 기능 중 일부도 해당 서버와 연결된 단말 또는 디바이스에서 수행될 수도 있다.

[0021] 본 명세서 내에 기재된 구성요소에 의해 실현되는 기능은 해당 기재된 기능을 실현하도록 프로그램된 범용 프로세서, 특정 용도 프로세서, 집적회로, ASICs(Application Specific Integrated Circuits), CPU(Central Processing Unit), 회로 및/또는 이들 조합을 포함하는 프로세싱 회로(processing circuitry)에서 구현되어도 된다. 프로세서는 트랜지스터나 기타 회로를 포함하며, 회로 또는 프로세싱 회로 로 간주된다. 프로세서는 메모리에 저장된 프로그램을 실행하는 프로그램된 프로세서(programmed processor)여도 좋다.

[0022] 본 명세서에서, 회로, 부, 유닛, 수단은 기재된 기능을 실현하도록 프로그램된 하드웨어 또는 실행하는 하드웨어이다. 해당 하드웨어는 본 명세서에 개시된 모든 하드웨어 또는 해당 기재된 기능을 실현하도록 프로그램되거나 실행하는 것으로 알려진 임의의 하드웨어라도 무방하다.

[0023] 해당 하드웨어가 회로 타입이라고 간주되는 프로세서인 경우, 해당 회로, 해당 부, 수단 또는 유닛은 하드웨어와 해당 하드웨어 및 또는 프로세서를 구성하기 위해 이용되는 소프트웨어의 조합이다.

[0024] 이하 첨부된 도면을 참고하여 본 발명의 일 실시예를 상세히 설명하기로 한다.

[0025] 도 1내지 도 3은 본 발명의 일 실시예에 따른 음성 변환 방법을 설명하기 위한 흐름도이다.

[0026] 도 1을 참조하면, 음성 변환 장치에 의한 음성 변환 방법은 소스 음성 신호와 타겟 음성 신호를 수신하는 음성 신호 수신 단계(S100), 소스 음성 신호와 타겟 음성 신호를 인코더에 입력하고, 상기 인코더로부터 상기 소스 음성 신호와 상기 타겟 음성 신호 각각의 컨텐츠 벡터와 스타일 벡터가 출력되는 인코딩 단계(S110), 출력된 컨텐츠 벡터 및 스타일 벡터를 디코더에 입력하고, 디코더로부터 합성 음성 신호가 출력되는 디코딩 단계(S120)를 포함할 수 있다.

[0027] 음성 신호가 입력되면, 해당 신호는 시계열에 따라 프레임 윈도우로 분리될 수 있다. 분리된 프레임 윈도우는 음성 임베딩 레이어를 통과하여 벡터로 변환되고, 이후 각 벡터에는 위치 값을 나타내는 포지셔널 임베딩 정보가 추가된다. 이렇게 구성된 벡터들은 트랜스포머 인코더에 입력되어, 음성의 특징을 더욱 정교하게 파악하고 처리할 수 있게 된다.

[0028] 상기 임베딩 벡터로부터 공유 벡터를 출력하는 트랜스포머 인코더, 상기 공유 벡터로부터 상기 컨텐츠 벡터를 출력하는 컨텐츠 인코더 및 상기 공유 벡터로부터 상기 스타일 벡터를 출력하는 스타일 인코더를 포함할 수 있다.

[0029] 그리고 합성 음성 신호는 소스 음성 신호의 컨텐츠 벡터와 타겟 음성 신호의 스타일 벡터가 합성된 음성 신호일 수 있다.

[0030] 음성 변환 장치는 음성 변환 모델을 이용해 입력 음성 신호에 대해 음성 변환을 수행하는 장치를 의미할 수 있다.

[0031] 음성 변환 장치 및 후술할 음성 변환 모델 학습 장치는 서버, 단말, 또는 이들이 결합된 형태일 수 있다. 여기서 단말은 각각 메모리(memory), 프로세서(processor)를 구비함으로써 연산 처리 능력을 갖춘 장치를 통칭하는 것이다. 예를 들어, 퍼스널 컴퓨터(personal computer), 핸드헬드 컴퓨터(handheld computer), PDA(personal digital assistant), 휴대폰, 스마트 기기, 태블릿(tablet) 등이 있다.

[0032] 서버는 복수개의 모듈(module)이 저장되어 있는 메모리, 그리고 메모리에 연결되어 있고 복수개의 모듈에 반응하며, 단말에 제공하는 서비스 정보 또는 서비스 정보를 제어하는 액션(action) 정보를 처리하는 프로세서, 통신 수단, 그리고 UI(user interface) 표시 수단을 포함할 수 있다.

[0033] 메모리는 정보를 저장하는 장치로, 고속 랜덤 액세스 메모리(high-speed random access memory, 자기 디스크 저장 장치, 플래시 메모리 장치, 기타 비휘발성 고체 상태 메모리 장치(non-volatile solid-state memory device) 등의 비휘발성 메모리 등 다양한 종류의 메모리를 포함할 수 있다.

[0034] 도 2를 통해, 음성 변환 장치에 의해 음성 변환 방법이 수행되는 과정을 보다 구체적으로 알 수 있다.

[0035] 도 2를 참조하면, 음성 변환 장치에 소스 음성 신호(201)와 타겟 음성 신호(202)가 입력되는 것을 알 수 있다. 여기서, 소스 음성 신호(201)와 타겟 음성 신호(202)는 음성 변환 장치의 사용자에 입력되는 음성 신호로서, 사용자가 직접 발화한 것을 녹음한 것이거나 다른 사람이 발화한 것을 녹음한 것일 수도 있다. 소스 음성 신호

(201)와 타겟 음성 신호(202)는 음성 데이터가 시계열적으로 저장된 파일이거나 스트리밍 형태로 실시간 입력될 수도 있다.

[0036] 소스 음성 신호(201)와 타겟 음성 신호(202)는 음성 신호 내 발화자가 발화하는 콘텐츠 내용에 대한 언어적 특징을 포함하는 컨텐츠 정보와 발화자의 특징, 억양, 성량, 패턴 등 비언어적 특징을 포함하는 정보인 스타일 정보를 포함할 수 있다.

[0037] 본 발명의 음성 변환 모델은 소스 음성 신호(201)와 타겟 음성 신호(202)에 기초하여 합성 음성 신호(209)를 생성할 수 있으며, 합성 음성 신호(209)는 소스 음성 신호(201)에 포함된 컨텐츠 정보에 타겟 음성 신호(202)의 스타일 정보가 합성된 음성 신호를 의미할 수 있다.

[0038] 본 발명에서 음성 변환 모델은 인코더(203)와 디코더(208)를 포함할 수 있다.

[0039] 인코더(203)는 소스 음성 신호(201)로부터 소스 컨텐츠 벡터(204)와 소스 스타일 벡터(205)를 출력하고, 타겟 음성 신호(202)로부터 타겟 컨텐츠 벡터(206)와 타겟 스타일 벡터(207)를 출력할 수 있다. 인코더(203)는 입력되는 음성 신호의 각 부분 간의 관계를 파악하고, 이를 통해 고차원의 벡터값을 생성할 수 있다. 이러한 벡터값은 음성의 특징을 포괄적으로 나타내며, 컨텐츠와 스타일을 분리하기 위한 초기 단계로 활용될 수 있다. 인코더(203)의 보다 구체적인 구성은 후술하기로 한다.

[0040] 컨텐츠 벡터는 대응되는 음성 신호의 언어적 내용, 즉 발화자가 말하는 내용이 무엇인 지에 대한 정보를 담고 있다. 이는 음성의 의미론적 요소를 추출하는 과정으로, 변환할 음성의 실질적인 내용을 포착하고 보존하는 역할을 한다. 이하에서, 소스 음성 신호(201)로부터 도출되는 컨텐츠 벡터는 소스 컨텐츠 벡터(204)로 정의하고, 타겟 음성 신호(202)로부터 도출되는 컨텐츠 벡터는 타겟 컨텐츠 벡터(206)로 정의한다.

[0041] 스타일 벡터는 음성 신호 내 음성의 프로소디(prosody)적 특성을 나타내며, 여기에는 억양, 음조, 리듬과 같은 요소들이 포함될 수 있다. 이는 변환하려는 대상 음성의 개성과 특성을 캡처하는데 중점을 둔다. 이하에서, 소스 음성 신호(201)로부터 도출되는 스타일 벡터는 소스 스타일 벡터(205)로 정의하고, 타겟 음성 신호(202)로부터 도출되는 스타일 벡터는 타겟 스타일 벡터(207)로 정의한다.

[0042] 인코더(203)로부터 출력되는 소스 컨텐츠 벡터(204), 소스 스타일 벡터(205), 타겟 컨텐츠 벡터(206) 및 타겟 스타일 벡터(207)는 디코더(208)에 입력될 수 있다.

[0043] 디코더(208)는 입력된 각각의 벡터값을 조합하여 최종적인 합성 음성 신호(209)를 생성할 수 있다. 디코더(208)는 음성의 최종 형태를 결정하는 단계로, 입력된 벡터들을 실제 청취 가능한 음성 신호로 변환하는 과정을 담당한다. 디코더(208)는 컨텐츠 벡터와 스타일 벡터를 적절히 합성하여 컨텐츠 벡터가 제공하는 언어적 내용과 스타일 벡터가 제공하는 음성의 특성을 결합하여 자연스럽고 일관된 음성 출력을 생성할 수 있다.

[0044] 디코더(208)는 내부의 신경망 구조에 의해 수행될 수 있고, 신경망의 가중치와 파라미터를 통해 컨텐츠 벡터와 스타일 벡터를 합성할 수 있다. 결합된 벡터는 디코더(208)의 신경망을 통해 처리되며 이러한 과정에서 신경망은 학습된 패턴을 바탕으로 합성 음성 신호를 생성할 수 있다. 이러한 신경망은 음성의 자연스러움 흐름과 리듬을 재현하는데 중점을 둔다.

[0045] 도 3은 본 발명의 실시예에 따른 음성 변환 모델의 인코더를 보다 상세히 설명하기 위한 구성도이다.

[0046] 도 3을 참조하면, 인코더는 음성 신호(301)가 입력되면 음성 신호(301)를 시계열에 따른 프레임 윈도우로 분리하고, 분리된 프레임 윈도우 별로 임베딩 벡터(303)를 출력하는 음성 임베딩 레이어(302)를 포함할 수 있다.

[0047] 음성 임베딩 레이어(302)는 입력되는 음성 신호(301)로부터 1DCNN(1-Dimensional Convolutional Neural Network)를 통해 임베딩을 수행할 수 있다. 음성 임베딩 레이어(302)를 통해 음성 신호(301)는 프레임 단위 또는 프레임 윈도우 단위로 임베딩 되고, 각 프레임 별 위치에 대한 포지셔널 임베딩 정보가 추가되어 임베딩 벡터(303)가 출력될 수 있다.

[0048] 음성 임베딩 레이어(302)는 상기한 1DCNN 뿐 아니라, 시계열적 데이터를 처리할 수 있는 다른 CNN(Convolutional Neural Network)이나 LSTM(Long Short Term Memory)에 반하려 임베딩 벡터(303)를 출력할 수도 있다.

[0049] 임베딩 벡터(303)는 트랜스포머 인코더(304)에 입력되어 공유 벡터(305)로 출력될 수 있다.

[0050] 트랜스포머 인코더(304)는 입력된 음성 신호(301)에서 중요한 정보를 추출하는 역할을 할 수 있다. 이러한 과

정에서 트랜스포머 인코더(304)는 자기 주의(self-attention) 메커니즘을 사용하여 음성 신호(301)의 다양한 특징과 패턴을 인식하고 이를 통해 고차원의 벡터값을 생성할 수 있다. 트랜스포머 인코더(304)는 주로 임베딩 벡터에 대해 적어도 일부를 마스킹(masking)하는 자기 지도 학습(Self Supervised Learning)을 통해 학습될 수 있다. 트랜스포머 인코더(304)가 사전 학습을 통해 공유 벡터(305)를 출력하면, 발화자 식별 번호와 컨텐츠 식별 번호를 이용하여 컨텐츠 인코더(306), 스타일 인코더(310), 컨텐츠 분류기(308) 및 발화자 분류기(312)를 다운스트림 태스크(Down-stream task)로 학습할 수 있다.

[0051] 트랜스포머 인코더(304)로부터 생성되는 공유 벡터(305)는 음성 신호(301)의 중요한 특징을 압축적으로 표현할 수 있다.

[0052] 컨텐츠 인코더(306)는 공유 벡터(305) 내의 정보 중에서 언어적 컨텐츠와 관련된 부분을 식별하여 컨텐츠 벡터(307)로 변환할 수 있다. 이러한 과정에서, 컨텐츠 인코더(306)는 발화의 의미, 구문적 구조, 단어의 선택 등 언어적 요소에 중점을 두고 정보를 추출할 수 있다.

[0053] 스타일 인코더(310)는 공유 벡터(305) 내의 정보 중에서 음성 신호의 스타일과 관련된 정보를 추출하여 스타일 벡터(311)로 변환할 수 있다. 이러한 과정에서, 스타일 인코더(310)는 발화의 톤, 강조 패턴, 발음 속도와 같은 음성의 비언어적 요소에 중점을 두어 정보를 추출할 수 있다.

[0054] 컨텐츠 벡터(307)는 컨텐츠 분류기(308)에 입력되어 컨텐츠 예측값(309)으로 출력될 수 있다. 컨텐츠 인코더(306)로부터 생성된 컨텐츠 벡터는 구체적인 음성 내용을 담고 있다. 컨텐츠 분류기(308)는 컨텐츠 벡터(307)를 분석하여, 음성 신호(301) 내의 특정 내용이나 대화 주제 등을 식별하고, 이에 대한 식별 번호나 태그를 컨텐츠 예측값(309)으로 출력할 수 있다.

[0055] 스타일 벡터(311)는 발화자 분류기(312)에 입력되어 발화자 예측값(313)으로 출력될 수 있다. 발화자 분류기(312)는 음성 신호(301) 내의 발화자의 신원을 확인할 수 있다. 스타일 인코더(310)로부터 생성된 스타일 벡터(311)는 발화자의 음성 스타일과 관련된 정보를 담고 있다. 발화자 분류기(312)는 스타일 벡터(311)를 분석하여 발화자의 신원이나 특징을 식별하고, 이를 통해 해당 발화자의 신원을 발화자 예측값(313)으로 출력할 수 있다.

[0056] 다시 도 1로 돌아오면, 본 발명의 실시예에 따른 음성 변환 방법은 인코더와 디코더를 학습시키는 학습 단계(S130)를 더 포함할 수 있다.

[0057] 학습 단계(S130)는 학습 소스 음성 신호와 학습 타겟 음성 신호가 상기 인코더에 입력되는 입력 단계, 상기 인코더가 상기 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩 단계, 상기 인코더가 상기 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩 단계, 상기 디코더가 상기 소스 컨텐츠 벡터, 상기 소스 스타일 벡터, 상기 타겟 컨텐츠 벡터 및 상기 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩 단계 및 상기 학습 합성 음성 신호에 손실 함수를 적용하고, 상기 손실 함수에 의한 손실이 적어지도록 상기 디코더를 튜닝하는 튜닝 단계; 를 포함할 수 있다.

[0058] 도 1에서, 학습 단계(S130)는 음성 변환 장치에 의해 수행되는 것으로 예시되었으나, 음성 변환 모델을 학습시키는 단계는 음성 변환 장치가 아닌 별도의 장치(예를 들어, 음성 변환 모델 학습 장치)에서 수행될 수 있다. 즉, 본 발명의 다른 실시예에서 음성 변환 장치는 음성 변환 모델 학습 장치에 의해 학습이 완료된 음성 변환 모델을 가져와 사용할 수 있다. 음성 변환 모델을 학습시키는 학습 단계(S130)에 대해선, 이하 도 4 내지 도 7을 통해 보다 구체적으로 설명하기로 한다.

[0059] 도 4와 도 5는 본 발명의 일 실시예에 따른 음성 변환 모델 학습 방법을 설명하기 위한 흐름도이다.

[0060] 도4를 참조하면, 음성 변환 모델 학습 방법은 학습 소스 음성 신호와 학습 타겟 음성 신호가 인코더에 입력되는 학습 음성 신호 입력 단계(S200), 인코더가 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩 단계(S210), 인코더가 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩 단계(S220), 디코더가 소스 컨텐츠 벡터, 소스 스타일 벡터, 타겟 컨텐츠 벡터 및 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩 단계(S230) 및 학습 합성 음성 신호에 손실 함수를 적용하여 상기 디코더를 튜닝하는 튜닝 단계(S240)를 포함할 수 있다.

[0061] 음성 변환 모델 학습 방법은 음성 변환 모델 학습 장치에 의해 수행될 수 있으나, 상기한 바와 같이 음성 변환 모델 학습 방법은 도 1을 통해 설명한 음성 변환 장치에 의해 수행될 수도 있다.

[0062] 여기서 학습 소스 음성 신호의 학습 타겟 음성 신호는 같은 발화자에 의해 생성된 음성 신호일 수 있다.

[0063] 그리고 손실 함수는 상기 학습 소스 음성 신호 또는 상기 학습 타겟 음성 신호와 상기 학습 합성 음성 신호 간의 차이를 계산하는 신호 복원 손실 함수를 포함할 수 있다.

[0064] 도 5는 본 발명의 일 실시예에 따른 음성 변환 모델 학습 방법 중 인코더를 학습시키는 방법을 설명하기 위한 도면이다.

[0065] 도 5를 참조하면, N번 발화자의 M번 컨텐츠 음성 신호(501)가 인코더(502)에 입력되는 것을 알 수 있다. 상기 음성 신호(501)는 N번으로 예시되어 식별되는 발화자(511)의 발화 스타일을 적용하여, M번으로 예시되어 식별되는 컨텐츠(506)를 발화한 것을 녹음한 신호일 수 있다.

[0066] 인코더(502)는 음성 신호(501)로부터 컨텐츠 벡터(503)와 스타일 벡터(508)를 출력할 수 있다. 컨텐츠 벡터(503)는 컨텐츠 분류기(504)에 입력되어 컨텐츠 예측값(505)으로 출력될 수 있다. 그리고 스타일 벡터(508)는 발화자 분류기(509)에 입력되어 발화자 예측값(510)으로 출력될 수 있다.

[0067] 이후, 음성 변환 모델 학습 장치는 컨텐츠 판별 손실 함수(507)를 통해 컨텐츠 분류기(504)로부터 출력된 컨텐츠 예측값(505)과 음성 신호(501)의 M번 컨텐츠(506)의 차이를 계산할 수 있다. 그리고 음성 변환 모델 학습 장치는 컨텐츠 판별 손실 함수(507)에 의해 도출된 값이 적어지도록 상기 컨텐츠 분류기(504)를 학습시킬 수 있다.

[0068] 마찬가지로, 음성 변환 모델 학습 장치는 발화자 판별 손실 함수(512)를 통해 발화자 분류기(509)로부터 출력된 발화자 예측값(510)과 음성 신호(501)의 N번 발화자(511)의 차이를 계산할 수 있다. 그리고 음성 변환 모델 학습 장치는 발화자 판별 손실 함수(512)에 의해 도출된 값이 적어지도록 상기 발화자 분류기(509)를 학습시킬 수 있다.

[0069] 도 6은 본 발명의 실시예에 따른 음성 변환 모델 학습 방법을 설명하기 위한 흐름도이다. 도 6의 음성 변환 모델 학습 방법은 도 4의 음성 변환 모델 학습 방법과 비교하여 복원 인코딩 단계(S340)가 더 추가된다.

[0070] 도 6과 도 7을 설명함에 있어서, 앞선 도면과 중복되는 구성이나 기능, 효과에 대한 설명은 생략하기로 한다.

[0071] 보다 구체적으로, 음성 변환 모델 학습 방법은 학습 소스 음성 신호와 학습 타겟 음성 신호가 인코더에 입력되는 학습 음성 신호 입력 단계(S300), 인코더가 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩 단계(S310), 인코더가 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩 단계(S320), 디코더가 소스 컨텐츠 벡터, 소스 스타일 벡터, 타겟 컨텐츠 벡터 및 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩 단계(S330), 상기 학습 합성 음성 신호가 복원 인코더에 입력되고, 상기 복원 인코더가 상기 학습 합성 음성 신호로부터 복원 컨텐츠 벡터와 복원 스타일 벡터를 출력하는 복원 인코딩 단계(S340) 및 상기 학습 합성 음성 신호, 상기 복원 컨텐츠 벡터 및 상기 복원 스타일 벡터 중 적어도 하나 이상에 상기 손실 함수를 적용하여 상기 디코더를 튜닝하는 튜닝 단계(S350)를 포함할 수 있다.

[0072] 그리고 손실 함수는 소스 컨텐츠 벡터 또는 타겟 컨텐츠 벡터와 복원 컨텐츠 벡터 간의 차이를 계산하는 컨텐츠 복원 손실 함수, 소스 스타일 벡터 또는 타겟 스타일 벡터와 상기 복원 스타일 벡터 간의 차이를 계산하는 컨텐츠 복원 손실 함수를 포함할 수 있다.

[0073] 여기서, 디코더는 생성적 대립 신경망(Generative Adversarial Network)에 기반할 수 있다. 또한, 디코더는 디퓨전 모델(Diffusion model)에 기반할 수 있다.

[0074] 도 7은 도 6의 음성 변환 모델 학습 방법을 보다 상세히 설명하기 위한 도면이다.

[0075] 디코더(708)의 기본적인 동작은 입력된 컨텐츠 벡터와 스타일 벡터를 합성하여 합성 음성 신호를 생성하는 것이다. 이를 위해, 상기한 방식으로 기 학습된 인코더(703)를 사용할 수 있다.

[0076] 우선, 학습 소스 음성 신호(701)와 학습 타겟 음성 신호(702)가 인코더(703)에 입력될 수 있다. 이 때, 학습 소스 음성 신호(701)는 N번 발화자의 M번 컨텐츠를 포함하는 것으로 정의하고, 학습 타겟 음성 신호(702)는 N번 발화자의 P번 컨텐츠 음성 신호를 포함하는 것으로 정의할 수 있다.

[0077] 즉, 학습 소스 음성 신호(701)와 학습 타겟 음성 신호(702)는 같은 N번 스타일을 가진 발화자에 의해 생성된 음성 신호일 수 있다. 이하에서, 학습 소스 음성 신호(701)에 포함되는 N번 스타일과 학습 타겟 음성 신호(702)에

포함되는 N번 스타일을 서로 구분하기 위해, 각각 N1번과 N2번의 명칭을 사용한다. 그리고, 학습 소스 음성 신호(701)와 학습 타겟 음성 신호(702)의 컨텐츠는 각각 M번과 P번으로 서로 다를 수 있다.

[0078] 인코더(703)에 의해 출력되는 소스 컨텐츠 벡터(704)는 학습 소스 음성 신호(701)로부터 출력되는 벡터이며, M번 컨텐츠에 대한 벡터로 정의될 수 있다. 인코더(703)에 의해 출력되는 소스 스타일 벡터(705)는 학습 소스 음성 신호(701)로부터 출력되는 벡터이며, N1번 스타일에 대한 벡터로 정의될 수 있다. 인코더(703)에 의해 출력되는 타겟 컨텐츠 벡터(706)는 학습 타겟 음성 신호(702)로부터 출력되는 벡터이며, P번 컨텐츠에 대한 벡터로 정의될 수 있다. 인코더(703)에 의해 출력되는 타겟 스타일 벡터(707)는 학습 타겟 음성 신호(702)로부터 출력되는 벡터이며, N2번 스타일에 대한 벡터로 정의될 수 있다.

[0079] 디코더(708)는 상기 소스 컨텐츠 벡터(704), 소스 스타일 벡터(705), 타겟 컨텐츠 벡터(706) 및 타겟 스타일 벡터(707)를 입력받을 수 있고, 이 중 컨텐츠 벡터와 스타일 벡터를 합성하여 학습 합성 음성 신호(709)를 생성할 수 있다. 보다 구체적으로, 디코더(708)는 소스 컨텐츠 벡터(704)와 소스 스타일 벡터(705)를 합성한 제1 출력 신호, 소스 컨텐츠 벡터(704)와 타겟 스타일 벡터(707)를 합성한 제2 출력 신호, 타겟 컨텐츠 벡터(706)와 소스 스타일 벡터(705)를 합성한 제3 출력 신호, 타겟 컨텐츠 벡터(706)와 타겟 스타일 벡터(707)를 합성한 제4 출력 신호를 생성할 수 있다. 학습 합성 음성 신호(709)는 상기 제1 내지 제4 출력 신호를 포함할 수 있다.

[0080] 이후, 학습 합성 음성 신호(709) 즉, 상기 제1 내지 제4 출력 신호는 복원 인코더(711)에 입력되고, 복원 인코더(711)는 상기 제1 내지 제4 출력 신호로부터 복원 컨텐츠 벡터(712)와 복원 스타일 벡터(713)를 생성할 수 있다. 보다 구체적으로, 복원 인코더(711)는 제1 내지 제4 출력 신호로부터 각각 대응되는 제1 내지 제4 복원 신호를 출력할 수 있다.

[0081] 본 발명에서, 손실 함수는 학습 합성 음성 신호(709)와 학습 소스 음성 신호(701) 또는 학습 타겟 음성 신호(702)간의 차이를 계산하는 신호 복원 손실 함수(710)를 포함할 수 있다.

[0082] 또한, 손실 함수는 소스 컨텐츠 벡터(704) 또는 타겟 컨텐츠 벡터(706)와 복원 컨텐츠 벡터(712) 간의 차이를 계산하는 컨텐츠 복원 손실 함수(714)를 포함할 수 있다.

[0083] 또한, 손실 함수는 소스 스타일 벡터(705) 또는 타겟 스타일 벡터(707)와 상기 복원 스타일 벡터(713) 간의 차이를 계산하는 스타일 복원 손실 함수를 포함할 수 있다.

[0084] 이하에서, 수식을 통해 상기 손실 함수들을 보다 구체적으로 설명한다.

[0085] <수식 1>

[0086] $$\mathcal{L}_1 = ||\mathbf{x}_m^{(n)} - f_d(c_m, s_{n1})||^2$$

[0087] 수식 1에서 L_1은 제1 신호 복원 손실 함수를 의미할 수 있다. 그리고 $x_m^{(n)}$ 는 M번 컨텐츠와 N번 발화자를 포함하는 학습 소스 음성 신호를 의미하고, $f_d(c_m,s_{n1})$는 M번 컨텐츠 벡터와 N1번 스타일 벡터를 합성하여 디코더에 의해 출력된 제1 출력 신호를 의미할 수 있다.

[0088] <수식 2>

[0089] $$\mathcal{L}_2 = ||\mathbf{x}_m^{(n)} - f_d(c_m, s_{n2})||^2$$

[0090] 수식 2에서 L_2은 제2 신호 복원 손실 함수를 의미할 수 있다. 그리고 $x_m^{(n)}$ 는 M번 컨텐츠와 N번 발화자를 포함하는 학습 소스 음성 신호를 의미하고, $f_d(c_m,s_{n2})$는 M번 컨텐츠 벡터와 N2번 스타일 벡터를 합성하여 디코더에 의해 출력된 제2 출력 신호를 의미할 수 있다.

[0091] <수식 3>

[0092] $$\mathcal{L}_3 = ||\mathbf{x}_p^{(n)} - f_d(c_p, s_{n1})||^2$$

[0093] 수식 3에서 L_3은 제3 신호 복원 손실 함수를 의미할 수 있다. 그리고 $x_p^{(n)}$ 는 P번 컨텐츠와 N번 발화자를 포함하는 학습 소스 음성 신호를 의미하고, $f_d(c_p,s_{n1})$는 P번 컨텐츠 벡터와 N1번 스타일 벡터를 합성하여 디코더에 의해 출력된 제3출력 신호를 의미할 수 있다.

[0094] <수식 4>

[0095] $$\mathcal{L}_4 = ||\mathbf{x}_p^{(n)} - f_d(c_p, s_{n2})||^2$$

[0096] 수식 4에서 L_4은 제4 신호 복원 손실 함수를 의미할 수 있다. 그리고 $x_p^{(n)}$ 는 P번 컨텐츠와 N번 발화자를 포함하는 학습 소스 음성 신호를 의미하고, $f_d(c_p,s_{n2})$는 P번 컨텐츠 벡터와 N2번 스타일 벡터를 합성하여 디코더에 의해 출력된 제1 출력 신호를 의미할 수 있다.

[0097] <수식 5>

[0098] $$\mathcal{L}_5 = ||c_m - {}^*c_m^{(1)}||^2$$

[0099] 수식 5에서 L_5은 제1 컨텐츠 복원 손실 함수를 의미할 수 있다. 그리고 c_m은 M번 컨텐츠 벡터를 의미하고, $*c_m^{(1)}$ 은 제1 출력 신호로부터 복원 인코더에 의해 복원된 복원 컨텐츠 벡터를 의미한다.

[0100] <수식 6>

[0101] $$\mathcal{L}_6 = ||c_m - {}^*c_m^{(2)}||^2$$

[0102] 수식 6에서 L_6은 제2 컨텐츠 복원 손실 함수를 의미할 수 있다. 그리고 c_m은 M번 컨텐츠 벡터를 의미하고, $*c_m^{(2)}$ 은 제2 출력 신호로부터 복원 인코더에 의해 복원된 복원 컨텐츠 벡터를 의미한다.

[0103] <수식 7>

[0104] $$\mathcal{L}_7 = ||c_p - {}^*c_p^{(1)}||^2$$

[0105] 수식 7에서 L_7은 제3 컨텐츠 복원 손실 함수를 의미할 수 있다. 그리고 c_p은 P번 컨텐츠 벡터를 의미하고, $*c_p^{(1)}$ 은 제3 출력 신호로부터 복원 인코더에 의해 복원된 복원 컨텐츠 벡터를 의미한다.

[0106] <수식 8>

[0107] $$\mathcal{L}_8 = ||c_p - {}^*c_p^{(2)}||^2$$

[0108] 수식 8에서 L_8은 제4 컨텐츠 복원 손실 함수를 의미할 수 있다. 그리고 c_p은 P번 컨텐츠 벡터를 의미하고, $*c_p^{(2)}$ 은 제4 출력 신호로부터 복원 인코더에 의해 복원된 복원 컨텐츠 벡터를 의미한다.

[0109] <수식 9>

[0110] $$\mathcal{L}_9 = ||s_{n1} - {}^{*}s_{n1}^{(1)}||^2$$

[0111] 수식 9에서 L_9은 제1 스타일 복원 손실 함수를 의미할 수 있다. 그리고 s_{n1}은 N1번 스타일 벡터를 의미하고, ${}^{*}s_{n1}^{(1)}$ 은 제1 출력 신호로부터 복원 인코더에 의해 복원된 복원 스타일 벡터를 의미한다.

[0112] <수식 10>

[0113] $$\mathcal{L}_{10} = ||s_{n1} - {}^{*}s_{n1}^{(2)}||^2$$

[0114] 수식 10에서 L_{10}은 제2 스타일 복원 손실 함수를 의미할 수 있다. 그리고 s_{n1}은 N1번 스타일 벡터를 의미하고, ${}^{*}s_{n1}^{(2)}$ 은 제3 출력 신호로부터 복원 인코더에 의해 복원된 복원 스타일 벡터를 의미한다.

[0115] <수식 11>

[0116] $$\mathcal{L}_{11} = ||s_{n2} - {}^{*}s_{n2}^{(1)}||^2$$

[0117] 수식 11에서 L_{11}은 제3 스타일 복원 손실 함수를 의미할 수 있다. 그리고 s_{n2}은 N2번 스타일 벡터를 의미하고, ${}^{*}s_{n2}^{(1)}$ 은 제2 출력 신호로부터 복원 인코더에 의해 복원된 복원 스타일 벡터를 의미한다.

[0118] <수식 12>

[0119] $$\mathcal{L}_{12} = ||s_{n2} - {}^{*}s_{n2}^{(2)}||^2$$

[0120] 수식 12에서 L_{12}은 제4 스타일 복원 손실 함수를 의미할 수 있다. 그리고 s_{n2}은 N2번 스타일 벡터를 의미하고, ${}^{*}s_{n2}^{(2)}$ 은 제4 출력 신호로부터 복원 인코더에 의해 복원된 복원 스타일 벡터를 의미한다.

[0121] <수식 13>

$$
\begin{aligned}
{}^{*}c_m^{(1)}, {}^{*}s_{n1}^{(1)} &= f_e(f_d(c_m, s_{n1})) \\
{}^{*}c_m^{(2)}, {}^{*}s_{n2}^{(1)} &= f_e(f_d(c_m, s_{n2})) \\
{}^{*}c_p^{(1)}, {}^{*}s_{n1}^{(2)} &= f_e(f_d(c_p, s_{n1})) \\
{}^{*}c_p^{(2)}, {}^{*}s_{n2}^{(2)} &= f_e(f_d(c_p, s_{n2}))
\end{aligned}
$$

[0122]

[0123] 수식 13은 상기 수식 5 내지 수식 12에서 사용된 복원 컨텐츠 벡터와 복원 스타일 벡터를 도출하기 위한 수식이다. f_e는 복원 인코더로부터 출력되는 결과를 의미하고, f_d는 디코더에 의해 출력되는 결과를 의미할 수 있다.

[0124] 수식 13을 통해, 제1 복원 신호 $*c_m^{(1)}$, $*s_{n1}^{(1)}$은 제1 출력 신호 c_m, s_{n1}에 기초하여 생성되고, 제2 복원 신호 $*c_m^{(2)}$, $*s_{n1}^{(1)}$은 제2 출력 신호 c_m, s_{n2}에 기초하여 생성되고, 제3 복원 신호 $*c_p^{(1)}$, $*s_{n1}^{(2)}$은 제3 출력 신호 c_p, s_{n1}에 기초하여 생성되고, 제4 복원 신호 $*c_p^{(2)}$, $*s_{n1}^{(2)}$은 제2 출력 신호 c_p, s_{n2}에 기초하여 생성되는 것을 알 수 있다.

[0125] 상기 디코더는 생성적 대립 신경망(Generative Adversarial Network)에 기반할 수 있다. 디코더가 생성적 대립 신경망에 기반하는 경우, 수식 14를 통해 디코더의 손실 함수가 도출될 수 있다.

[0126] <수식 14>

$$
\min_{f_d} \max_{D} V(D, f_d) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(f_d(z, c, s))]
$$

[0127]

[0128] 수식 14에서 f_d는 디코더를 의미한다. D는 디코더의 아웃풋과 실제 음성 신호를 구분하는 판별 모델(Discriminator), z는 가우시안 노이즈, c와 s는 각각 컨텐츠 벡터와 스타일 벡터를 의미한다.

[0129] 디코더는 입력되는 음성 신호 x에 대해서, 판별 모델 D는 최대한 높은 확률을 출력하도록 학습될 수 있다. 반대로, f_d의 출력인 가짜 신호에 대해서 판별 모델 D는 최대한 낮은 확률을 출력하도록 학습될 수 있다.

[0130] 즉, 디코더f_d는 판별 모델 D의 확률값이 높아지도록, 즉 판별 모델D를 속이는 방향으로 학습될 수 있다.

[0131] 상기 디코더는 디퓨전 모델(Diffusion model)에 기반할 수 있다. 디코더가 디퓨전 모델에 기반하는 경우, 수식 15를 통해 디코더의 손실 함수가 도출될 수 있다.

[0132] <수식 15>

$$
L_{t-1} = D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||f_d(\mathbf{x}_{t-1}|\mathbf{x}_t, c, s))
$$

[0133]

[0134] 수식 15에서 f_d는 디코더를 의미한다. t는 디퓨전 단계(diffusion step), D_{KL}은 KL 발산(divergence), q는 디퓨전 과정(diffusion process)을 의미한다.

[0135] 노이즈 디퓨전 모형에 의거하여, 원본 신호 x_0가 주어졌을 때 시간 단계(time-step)의 t번째 신호 x_t는 다음 수식 16과 같이 정의될 수 있다.

[0136] <수식 16>

[0137] $$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t I))$$

[0138] 수식 16에서 t가 0에 가까울수록 분산 보존 요인(variance preserving factor) $\bar{\alpha}_t$의 크기가 커지며 $\bar{\alpha}_t$는 x_0과 유사해진다. 반대로, t가 무한대로 갈수록 $\bar{\alpha}_t$는 0에 수렴하며, x_t는 순수 가우시안 노이즈(pure gaussian noise)에 가까워진다. 노이즈 확산 프로세스를 거슬러 올라가면, 노이즈로부터 원본 신호 x_0을 복원할 수 있는데 이러한 역방향 확산 프로세스 $q(x_{t-1}|x_t)$는 계산 불가능하다. 대신, $q(x_{t-1}|x_t)$가 정규분포를 따른다는 성질을 이용하여 증거 하한(Evidence Lower Bound, ELBO)를 계산할 수 있고, 상기 증거 하한을 전개하여, 상수항을 제거하면 위에 기술한 L_{t-1}을 얻을 수 있다.

[0139] 다시 수식 15로 돌아와, 모든 시간 단계 t($0 \leq t \leq T$, 일반적으로 T=1000)에 대하여, L_{t-1}이 최소가 되는 방향으로 디코더 f_d를 학습할 수 있다.

[0140] 도 8은 본 발명의 일 실시예에 따른 음성 변환 장치를 설명하기 위한 구성도이다.

[0141] 도 8을 설명함에 있어서, 앞선 음성 변환 방법에서 설명된 내용과 중복되는 구성, 기능 및 효과는 기재를 생략한다.

[0142] 도 8을 참조하면, 음성 변환 장치(800)는 소스 음성 신호와 타겟 음성 신호를 수신하는 음성 신호 수신부(810), 소스 음성 신호와 타겟 음성 신호를 인코더에 입력하고, 인코더로부터 소스 음성 신호와 타겟 음성 신호 각각의 컨텐츠 벡터와 스타일 벡터가 출력되는 인코딩부(820), 출력된 컨텐츠 벡터 및 스타일 벡터를 디코더에 입력하고, 디코더로부터 합성 음성 신호가 출력되는 디코딩부(830)를 포함할 수 있다.

[0143] 그리고 음성 변환 장치(800)는 디코더를 학습시키는 학습부(840)를 포함할 수 있다.

[0144] 도 9는 본 발명의 일 실시예에 따른 음성 변환 모델 학습 장치를 설명하기 위한 구성도이다.

[0145] 도 9를 설명함에 있어서, 앞선 음성 변환 모델 학습 방법에서 설명된 내용과 중복되는 구성, 기능 및 효과는 기재를 생략한다.

[0146] 음성 변환 모델 학습 장치(900)는 학습 소스 음성 신호와 학습 타겟 음성 신호가 인코더에 입력되는 학습 음성 신호 입력부(910), 인코더가 학습 소스 음성 신호로부터 소스 컨텐츠 벡터와 소스 스타일 벡터를 출력하는 학습 소스 음성 신호 인코딩부(920), 인코더가 학습 타겟 음성 신호로부터 타겟 컨텐츠 벡터와 타겟 스타일 벡터를 출력하는 학습 타겟 음성 신호 인코딩부(930), 디코더가 소스 컨텐츠 벡터, 소스 스타일 벡터, 타겟 컨텐츠 벡터 및 타겟 스타일 벡터에 기초하여 학습 합성 음성 신호를 출력하는 학습 합성 음성 신호 디코딩부(940) 및 학습 합성 음성 신호에 손실 함수를 적용하여 디코더를 튜닝하는 튜닝부(950)를 포함할 수 있다.

[0147] 그리고, 음성 변환 모델 학습 장치(900)는 학습 합성 음성 신호가 복원 인코더에 입력되고, 상기 복원 인코더가 상기 학습 합성 음성 신호로부터 복원 컨텐츠 벡터와 복원 스타일 벡터를 출력하는 복원 인코딩부(960)를 포함할 수 있다.

[0148] 상기한 음성 변환 방법과 음성 변환 모델 학습 방법은 컴퓨터에 의해 실행되는 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램 또는 컴퓨터에 의해 실행 가능한 명령어들을 포함하는 기록 매체의 형태로도 구현될 수 있다. 또한, 상기한 음성 변환 방법과 음성 변환 모델 학습 방법은 컴퓨터에 의해 실행되는 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램의 형태로도 구현될 수 있다.

[0149] 컴퓨터 판독 가능 기록매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발

성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 기록매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0150] 상기 음성 변환 방법과 음성 변환 모델 학습 방법은 앞서 도 1 내지 도 9를 통해 설명된 실시예를 통해 더 추가적인 단계들로 더 분할되거나, 더 적은 단계들로 조합될 수 있다. 또한, 일부 단계는 필요에 따라 생략될 수도 있고, 단계 간의 순서가 전환될 수도 있다.

[0151] 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.

[0152] 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 것으로 해석되어야 한다.

부호의 설명

[0153] S100: 음성 신호 수신 단계

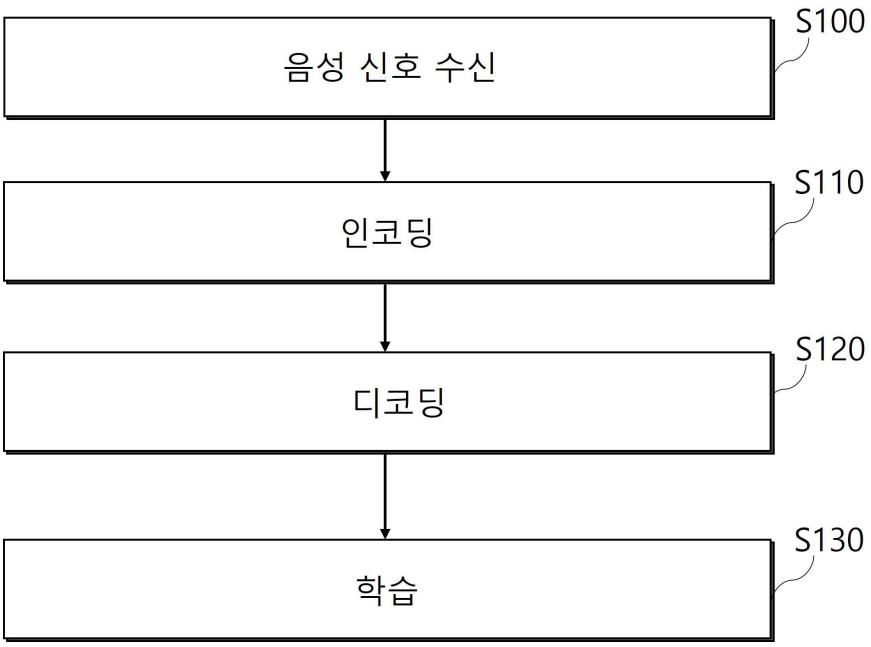
S200: 인코딩 단계

S300: 디코딩 단계

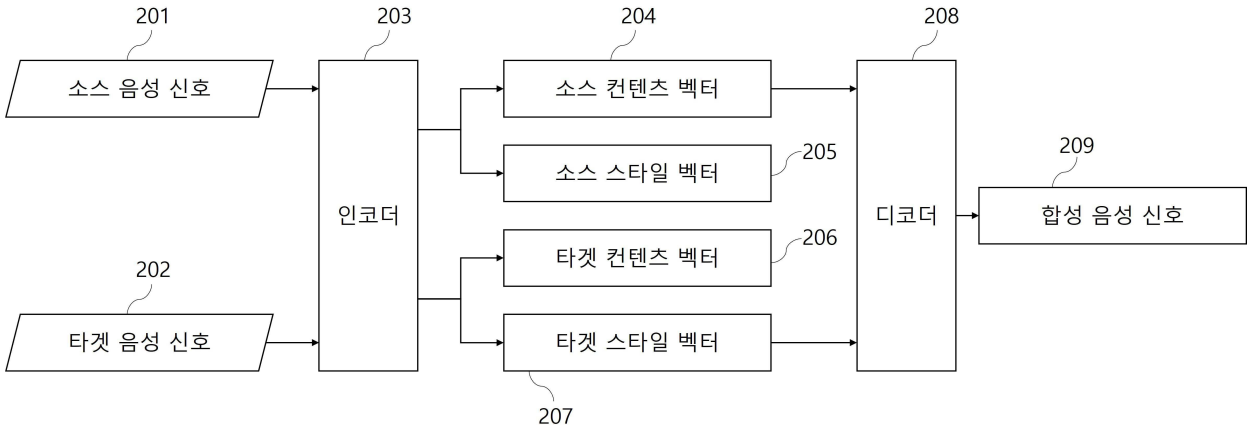
S400: 학습 단계

도면

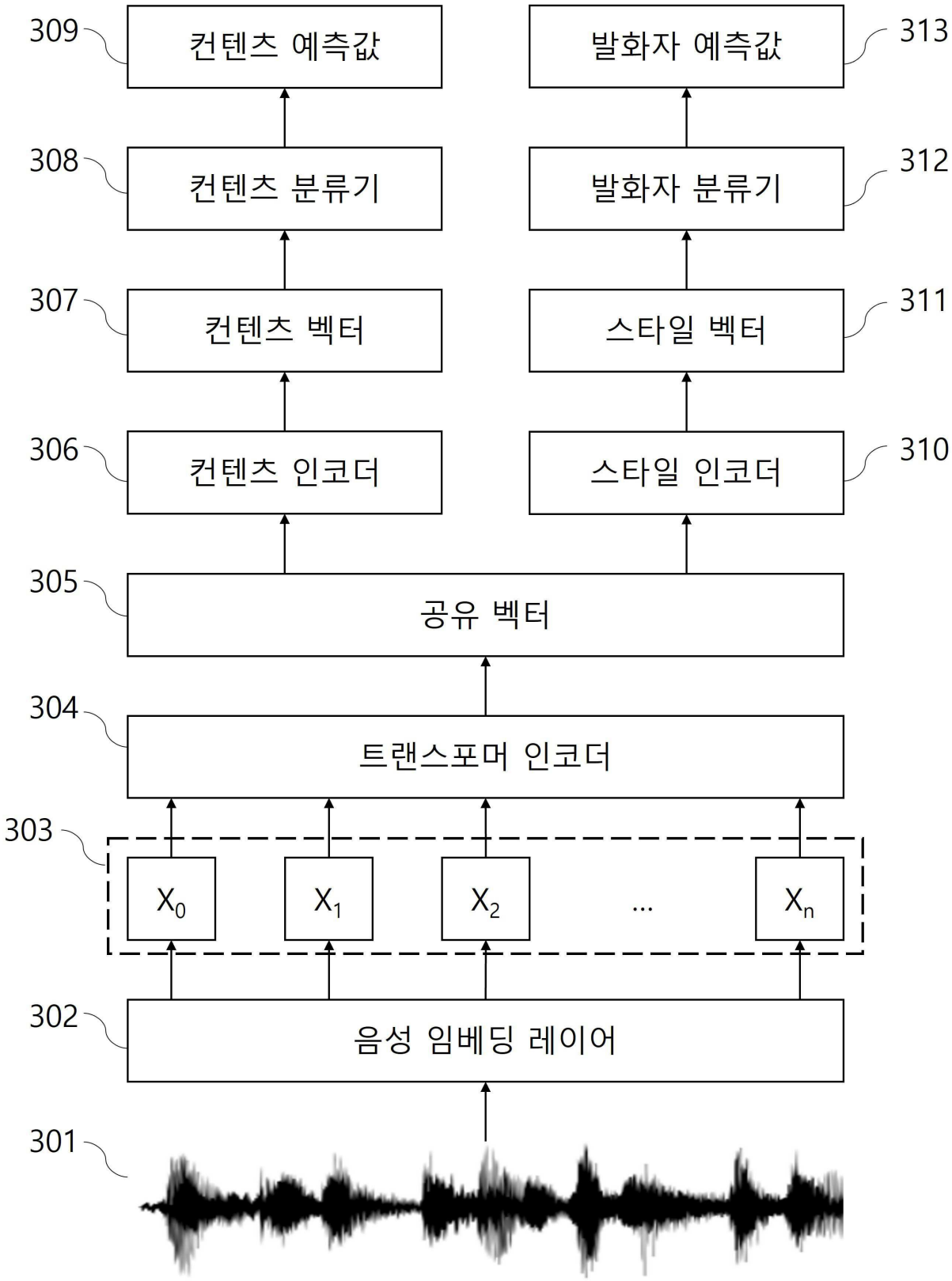
도면1



도면2


201
소스 음성 신호
202
타겟 음성 신호
203
인코더
204
소스 컨텐츠 벡터
205
소스 스타일 벡터
206
타겟 컨텐츠 벡터
207
타겟 스타일 벡터
208
디코더
209
합성 음성 신호

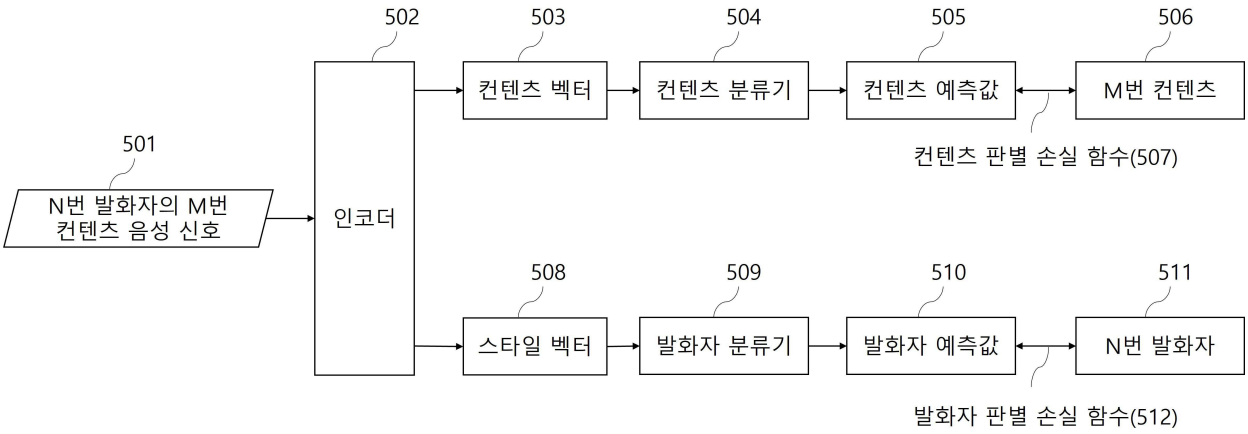
도면3


309
컨텐츠 예측값
313
발화자 예측값
308
컨텐츠 분류기
312
발화자 분류기
307
컨텐츠 벡터
311
스타일 벡터
306
컨텐츠 인코더
310
스타일 인코더
305
공유 벡터
304
트랜스포머 인코더
303
X_0
X_1
X_2
...
X_n
302
음성 임베딩 레이어
301

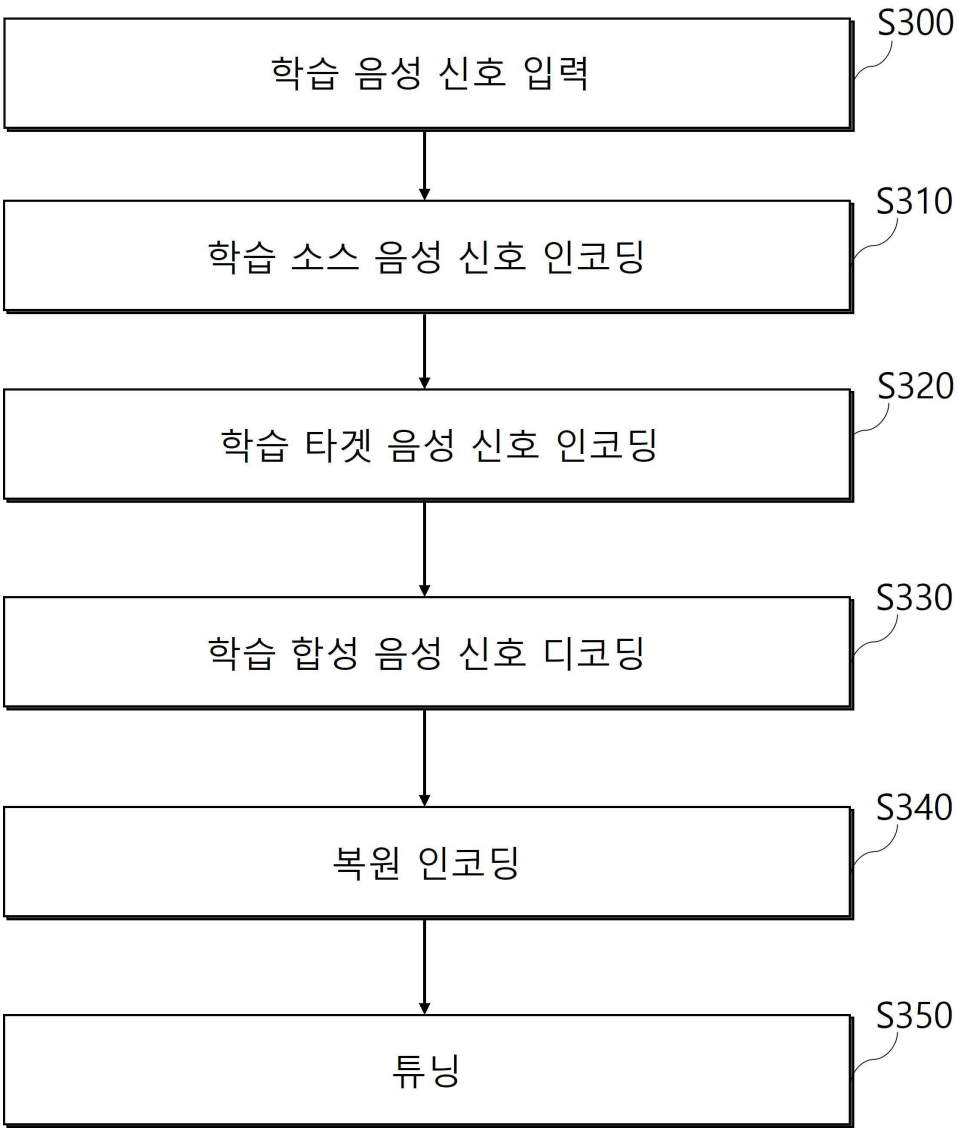
도면4


학습 음성 신호 입력
S200
학습 소스 음성 신호 인코딩
S210
학습 타겟 음성 신호 인코딩
S220
학습 합성 음성 신호 디코딩
S230
튜닝
S240

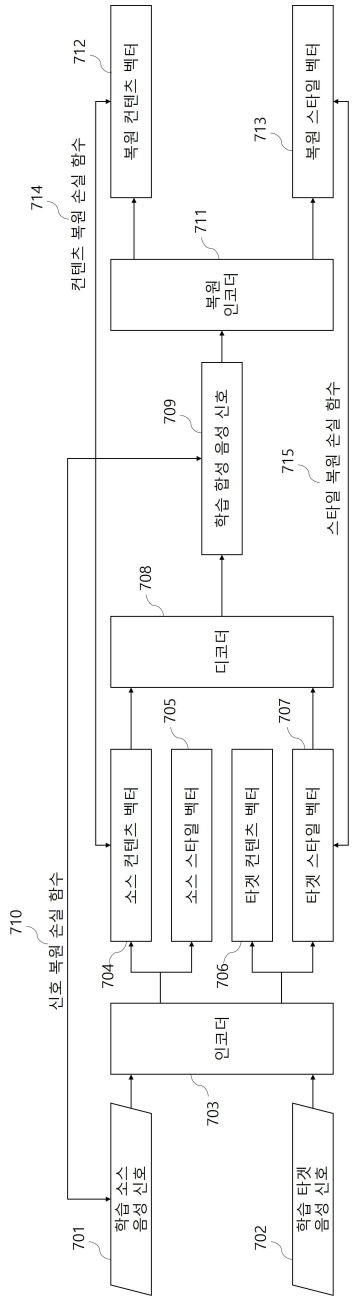
도면5


501
N번 발화자의 M번
컨텐츠 음성 신호
인코더
502
컨텐츠 벡터
503
컨텐츠 분류기
504
컨텐츠 예측값
505
M번 컨텐츠
506
컨텐츠 판별 손실 함수(507)
스타일 벡터
508
발화자 분류기
509
발화자 예측값
510
N번 발화자
511
발화자 판별 손실 함수(512)

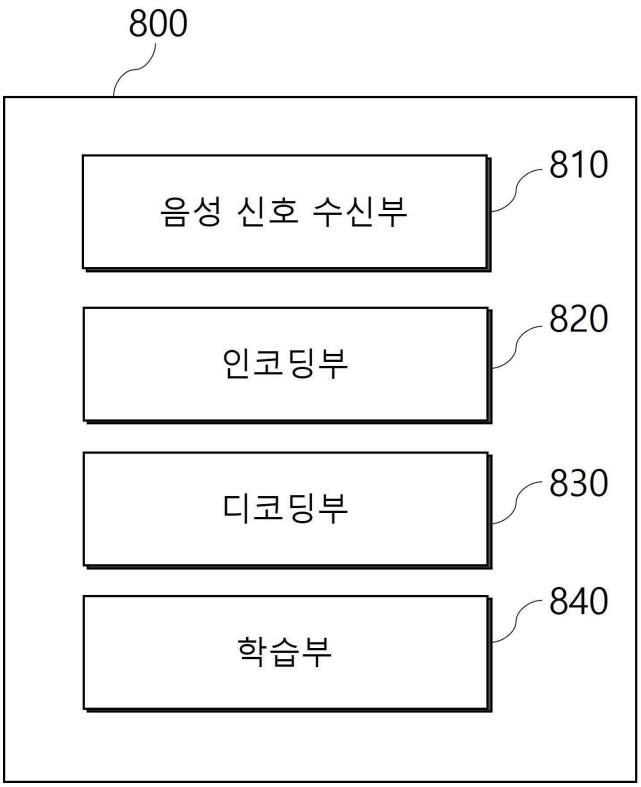
도면6


학습 음성 신호 입력
S300
학습 소스 음성 신호 인코딩
S310
학습 타겟 음성 신호 인코딩
S320
학습 합성 음성 신호 디코딩
S330
복원 인코딩
S340
튜닝
S350

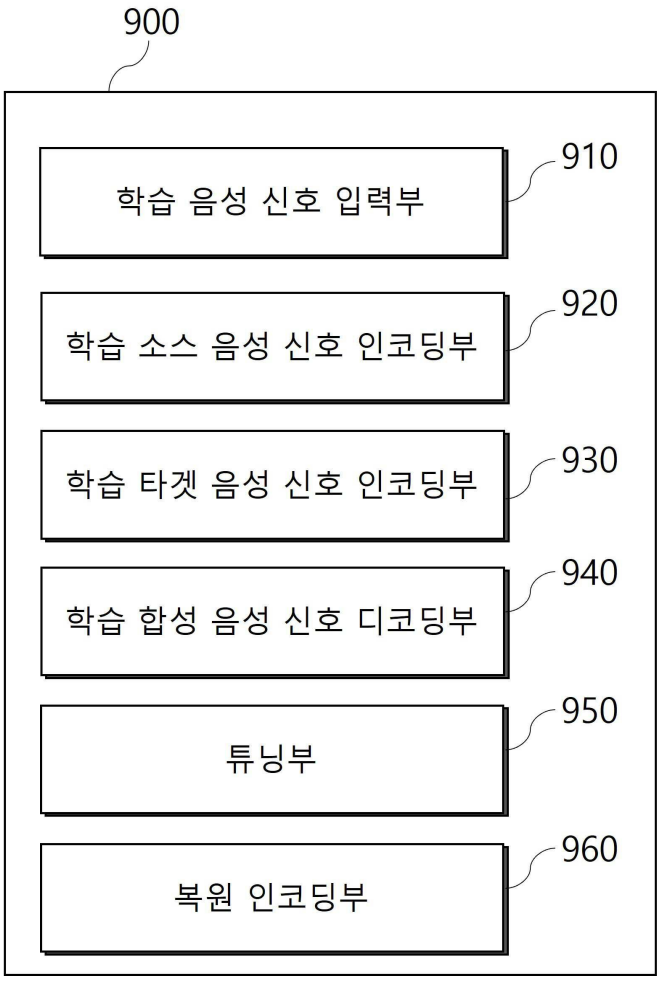
도면7


701
학습 소스 음성 신호
702
학습 타겟 음성 신호
703
인코더
704
소스 컨텐츠 벡터
705
소스 스타일 벡터
706
타겟 컨텐츠 벡터
707
타겟 스타일 벡터
708
디코더
709
학습 합성 음성 신호
710
신호 복원 손실 함수
711
복원 인코더
712
복원 컨텐츠 벡터
713
복원 스타일 벡터
714
컨텐츠 복원 손실 함수
715
스타일 복원 손실 함수

도면8



도면9


900
학습 음성 신호 입력부
910
학습 소스 음성 신호 인코딩부
920
학습 타겟 음성 신호 인코딩부
930
학습 합성 음성 신호 디코딩부
940
튜닝부
950
복원 인코딩부
960