(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0107703
(43) 공개일자 2022년08월02일

(51) 국제특허분류(Int. Cl.)
G10L 13/033 (2013.01) *G10L 17/02* (2013.01)
G10L 21/003 (2013.01) *G10L 21/013* (2013.01)
(52) CPC특허분류
G10L 13/033 (2013.01)
G10L 17/02 (2013.01)
(21) 출원번호 10-2021-0010669
(22) 출원일자 2021년01월26일
심사청구일자 2021년01월26일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
육동석
경기도 남양주시 천마산로 65, 104동 2004호(호평동, 호평 파라곤)
유인철
서울특별시 성북구 인촌로7마길 7, 201호(안암동1가)
장형필
서울특별시 강서구 초록마을로5길 40, 302호(화곡동, 삼성맨션)
(74) 대리인
송인호, 윤형근, 최관락

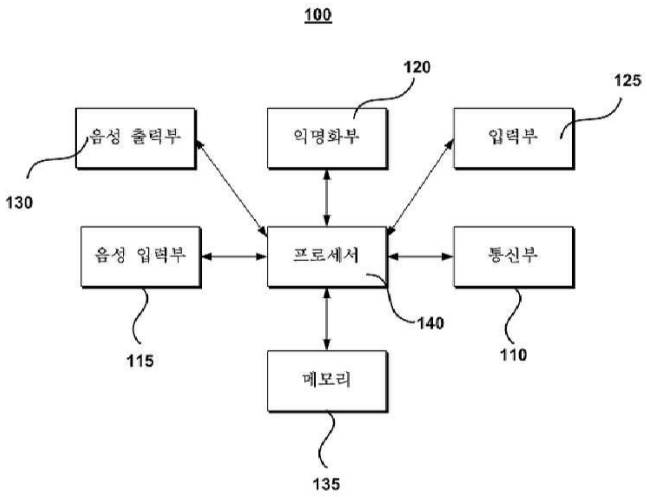
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **음성 익명화 방법 및 그 장치**

(57) 요 약

음성 익명화 방법 및 그 장치가 개시된다. 음성 익명화 방법은, 사용자의 음성 정보를 취득하는 단계; 및 상기 취득한 음성 정보를 음성 변환 모델에 적용하여 익명화하되, 익명의 화자 식별 벡터를 반영하여 상기 취득한 음성 정보를 변환함으로써 상기 취득한 음성 정보에서 발화 정보는 유지하며 화자 정보를 익명화하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 21/003 (2013.01)
G10L 2021/0135 (2013.01)

명 세 서

청구범위

청구항 1

사용자의 음성 정보를 취득하는 단계; 및

상기 취득한 음성 정보를 음성 변환 모델에 적용하여 익명화하되, 익명의 화자 식별 벡터를 반영하여 상기 취득한 음성 정보를 변환함으로써 상기 취득한 음성 정보에서 발화 정보는 유지하며 화자 정보를 익명화하는 단계를 포함하는 음성 익명화 방법.

청구항 2

제1 항에 있어서,

상기 익명의 화자 식별 벡터는,

대상 화자에 고정된 제1 값을 할당하고, 복수의 나머지 화자는 균등 분할된 제2 값을 할당하여 생성되되,

상기 제2 값은 상기 대상 화자에 할당되는 제1 값에 따라 달라지는 것을 특징으로 하는 음성 익명화 방법.

청구항 3

제2 항에 있어서,

상기 대상 화자와 상기 복수의 나머지 화자에 각각 할당되는 제1 값과 상기 제2 값들을 합산한 결과는 1이 되도록 상기 제1 값과 상기 제2 값이 결정되는 것을 특징으로 하는 음성 익명화 방법.

청구항 4

제1 항에 있어서,

상기 익명의 화자 식별 벡터는,

상기 대상 화자와 상기 복수의 나머지 화자에 균등 분할된 제1 값을 할당하여 생성되되,

상기 대상 화자에 할당되는 제1 값은 상기 복수의 나머지 화자에 할당되는 제1 값의 음수값이되, 상기 대상 화자와 상기 복수의 나머지 화자에 할당된 제1 값들을 합산한 결과가 1이 되도록 상기 제1 값이 결정되는 것을 특징으로 하는 음성 익명화 방법.

청구항 5

제1 항에 있어서,

상기 익명의 화자 식별 벡터는,

대상 화자와 복수의 나머지 화자 각각에 대한 i-벡터를 추출한 후 상기 추출된 i-벡터를 이용하여 상기 대상 화자와의 코사인 유사도를 각각 도출한 후 상기 코사인 유사도를 이용하여 상기 복수의 나머지 화자 중 적어도 일부에 값이 할당되는 것을 특징으로 하는 음성 익명화 방법.

청구항 6

제5 항에 있어서,

상기 익명의 화자 식별 벡터는,

상기 복수의 나머지 화자에 코사인 유사도의 역에 따라 정규화된 값을 상이하게 할당하되, 상기 복수의 나머지 화자에 할당된 값들을 합산한 결과는 1이 되도록 코사인 유사도의 역에 따라 상이하게 정규화된 값이 할당되는 것을 특징으로 하는 음성 익명화 방법.

청구항 7

제5 항에 있어서,

상기 익명의 화자 식별 벡터는,

상기 대상 화자와 코사인 유사도가 가장 먼 나머지 화자 또는 상기 대상 화자와 코사인 유사도가 가장 먼순으로 두명의 나머지 화자에 균등 분할된 값이 할당되되,

상기 익명의 화자 식별 벡터에 할당된 값들을 합산한 값은 1인 것을 특징으로 하는 음성 익명화 방법.

청구항 8

사용자의 음성 정보를 취득하는 음성 입력부; 및

상기 취득한 음성 정보를 음성 변환 모델에 적용하여 익명화하되, 익명의 화자 식별 벡터를 반영하여 상기 취득한 음성 정보를 변환함으로써 상기 취득한 음성 정보에서 발화 정보는 유지하며 화자 정보를 익명화하는 익명화부를 포함하는 음성 익명화 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 사용자의 발화 내용은 유지한 채로 신원을 특정할 수 있는 음성 특징을 제거하여 익명화할 수 있는 음성 익명화 방법 및 그 장치에 관한 것이다.

배 경 기 술

[0003] 음성 익명화는 발화 내용을 유지한 상태로, 말하는 사람의 신원을 특정하기 어렵도록 하는 기술이다. 음성 익명화와 비슷한 기술로는 제보자의 신원 보호를 위해 수행하는 음성 변조 기술이 있다. 그러나, 음성 변조 기술은 사용한 변조 알고리즘에 따라 역 필터(inverse filter) 등으로 원본 음성을 복원할 가능성이 있으며, 발화 내용 자체에도 손상이 가해져서 알아듣기 어려워지는 문제점이 있다. 이로 인해, 뉴스 등에서 음성 변조를 가한 경우 자막 등을 함께 표시하여 내용을 보완하고 있다.

[0004] 그러나, 음성 변환 기술은 발화 내용을 유지한 채로 말한 사람의 신원을 다른 사람의 것으로 변환하는 기술로, 예를 들어, 남성 목소리를 여성 목소리로 바꾸거나 노인 목소리를 어린이 목소리로 바꾸는 것이 가능하다.

[0005] 음성 인식과 같은 음성 기반의 인터페이스는 나날이 발전하고 있어 여러 분야에서 널리 활용되고 있다. 효과적인 음성 인식 시스템의 개발과, 자연어 이해 등의 연구를 위하여서는 대량의 음성 데이터를 수집할 필요가 있고, 특히나 다양한 사람들이 실제 환경에서 사용하는 자연스러운 대화체 문장들의 수집이 필수적이다. 하지만 음성 데이터는 말한 사람의 신원을 특정할 수 있을 성별, 연령등을 포함하는 발화 특성이 담긴 민감한 개인정보에 해당한다. 특히나 화자 인식과 같이 목소리를 이용한 사용자 인증에 직접적으로 사용될 수 있기에 취급과 관리에 각별한 주의가 필요하다.

[0006] 또한, 익명화 기술은 음성 인식, 자연어 이해 등에 사용할 수 있도록 발화한 내용 자체는 유지하면서 화자의 신원 특정할 수 있는 정보들을 제거하는 기술이나, 기존의 익명화 기술들은 일반 사용자들 입장에서 얼마나 자신

의 목소리에서 신원 정보가 사라졌는지를 확인할 수단을 제공하지 못하고 있다.

발명의 내용

해결하려는 과제

[0008] 본 발명은 사용자의 발화 내용은 유지한 채로 신원을 특정할 수 있는 음성 특징을 제거하여 익명화할 수 있는 음성 익명화 방법 및 그 장치를 제공하기 위한 것이다.

[0009] 또한, 본 발명은 소리 신호 수준에서 익명화를 적용하여 익명화가 완료된 소리를 사용자가 직접 들어보고 평가가 가능한 음성 익명화 방법 및 그 장치를 제공하기 위한 것이다.

[0010] 또한, 본 발명은 사용자들이 소리 재생 기기로 직접 듣고 평가가 가능하기 때문에, 여러 익명화 방식을 선택할 수 있으며, 선호하는 방식으로 익명화한 음성을 전송하도록 할 수 있는 음성 익명화 방법 및 그 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0012] 본 발명의 일 측면에 따르면, 사용자의 발화 내용은 유지한 채로 신원을 특정할 수 있는 음성 특징을 제거하여 익명화할 수 있는 음성 익명화 방법이 제공된다.

[0013] 본 발명의 일 실시예에 따르면, 사용자의 음성 정보를 취득하는 단계; 및 상기 취득한 음성 정보를 음성 변환 모델에 적용하여 익명화하되, 익명의 화자 식별 벡터를 반영하여 상기 취득한 음성 정보를 변환함으로써 상기 취득한 음성 정보에서 발화 정보는 유지하며 화자 정보를 익명화하는 단계를 포함하는 음성 익명화 방법이 제공될 수 있다. 상기 익명의 화자 식별 벡터는, 대상 화자에 고정된 제1 값을 할당하고, 복수의 나머지 화자는 균등 분할된 제2 값을 할당하여 생성되되, 상기 제2 값은 상기 대상 화자에 할당되는 제1 값에 따라 달라질 수 있다.

[0014] 상기 대상 화자와 상기 복수의 나머지 화자에 각각 할당되는 제1 값과 상기 제2 값들을 합산한 결과는 1이 되도록 상기 제1 값과 상기 제2 값이 결정될 수 있다.

[0015] 상기 익명의 화자 식별 벡터는, 상기 대상 화자와 상기 복수의 나머지 화자에 균등 분할된 제1 값을 할당하여 생성되되, 상기 대상 화자에 할당되는 제1 값은 상기 복수의 나머지 화자에 할당되는 제1 값의 음수값이되, 상기 대상 화자와 상기 복수의 나머지 화자에 할당된 제1 값들을 합산한 결과가 1이 되도록 상기 제1 값이 결정될 수 있다.

[0016] 상기 익명의 화자 식별 벡터는, 대상 화자와 복수의 나머지 화자 각각에 대한 i-벡터를 추출한 후 상기 추출된 i-벡터를 이용하여 상기 대상 화자와의 코사인 유사도를 각각 도출한 후 상기 코사인 유사도를 이용하여 상기 복수의 나머지 화자 중 적어도 일부에 값이 할당될 수 있다.

[0017] 상기 익명의 화자 식별 벡터는, 상기 복수의 나머지 화자에 코사인 유사도의 역에 따라 정규화된 값을 상이하게 할당하되, 상기 복수의 나머지 화자에 할당된 값들을 합산한 결과는 1이 되도록 코사인 유사도의 역에 따라 상이하게 정규화된 값이 할당될 수 있다.

[0018] 상기 익명의 화자 식별 벡터는, 상기 대상 화자와 코사인 유사도가 가장 먼 나머지 화자 또는 상기 대상 화자와 코사인 유사도가 가장 먼순으로 두명의 나머지 화자에 균등 분할된 값이 할당되되, 상기 익명의 화자 식별 벡터에 할당된 값들을 합산한 값은 1일 수 있다.

[0020] 본 발명의 다른 측면에 따르면, 사용자의 발화 내용은 유지한 채로 신원을 특정할 수 있는 음성 특징을 제거하여 익명화할 수 있는 장치가 제공된다.

[0021] 본 발명의 일 실시예에 따르면, 사용자의 음성 정보를 취득하는 음성 입력부; 및 상기 취득한 음성 정보를 음성 변환 모델에 적용하여 익명화하되, 익명의 화자 식별 벡터를 반영하여 상기 취득한 음성 정보를 변환함으로써 상기 취득한 음성 정보에서 발화 정보는 유지하며 화자 정보를 익명화하는 익명화부를 포함하는 음성 익명화 장

치가 제공될 수 있다.

발명의 효과

[0023] 본 발명의 일 실시예에 따른 음성 익명화 방법 및 그 장치를 제공함으로써, 사용자의 발화 내용은 유지한 채로 신원을 특정할 수 있는 음성 특징을 제거하여 익명화할 수 있는 이점이 있다.

[0024] 또한, 본 발명은 소리 신호 수준에서 익명화를 적용하여 익명화가 완료된 소리를 사용자가 직접 들어보고 평가가 가능한 이점이 있다.

[0025] 또한, 본 발명은 사용자들이 소리 재생 기기로 직접 듣고 평가가 가능하기 때문에, 여러 익명화 방식을 선택할 수 있으며, 선호하는 방식으로 익명화한 음성을 전송하도록 할 수 있는 이점도 있다.

도면의 간단한 설명

[0027] 도 1은 본 발명의 일 실시예에 따른 음성 익명화 장치의 구성을 개략적으로 도시한 블록도.

도 2는 본 발명의 일 실시예에 따른 음성 변환 모델의 세부 구조를 도시한 도면.

도 3은 본 발명의 일 실시예에 따른 제1 내지 제6 익명의 화자 식별 벡터를 설명하기 위해 도시한 도면.

도 4는 본 발명의 일 실시예에 따른 음성 익명화 방법을 나타낸 순서도.

도 5는 익명의 화자 식별 벡터에 대한 화자 식별 및 음성 인식 정확도를 나타낸 결과를 나타낸 도면.

발명을 실시하기 위한 구체적인 내용

[0028] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0029] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0031] 도 1은 본 발명의 일 실시예에 따른 음성 익명화 장치의 구성을 개략적으로 도시한 블록도이며, 도 2는 본 발명의 일 실시예에 따른 음성 변환 모델의 세부 구조를 도시한 도면이고, 도 3은 본 발명의 일 실시예에 따른 제1 내지 제6 익명의 화자 식별 벡터를 설명하기 위해 도시한 도면이다.

[0032] 도 1을 참조하면, 본 발명의 일 실시예에 따른 음성 익명화 장치(100)는 통신부(110), 음성 입력부(115), 익명화부(120), 입력부(125), 음성 출력부(130), 메모리(135) 및 프로세서(140)를 포함하여 구성된다.

[0033] 통신부(110)는 통신망을 통해 다른 장치들(예를 들어, 서버)와 데이터를 송수신하기 위한 수단이다.

[0034] 예를 들어, 통신부(110)는 프로세서(140)의 제어에 따라 익명화된 음성 정보를 서버로 전송할 수 있다. 프로세서(140)는 사용자에 의해 전송 동의가 완료된 경우 통신부(110)를 통해 익명화된 음성 정보를 서버로 전송하도록 제어할 수 있다.

[0035] 음성 입력부(115)는 사용자로부터 음성 정보를 입력받기 위한 수단이다. 예를 들어, 음성 입력부(115)는 마이크일 수 있다.

[0036] 익명화부(120)는 음성 변환 모델을 구비하되, 음성 변환 모델에 취득한 음성 정보를 적용하여 발화 내용을 유지한 상태로 화자 정보만 제거하여 익명화하기 위한 수단이다.

[0037] 본 발명의 일 실시예에 따르면, 익명화부(120)는 복수의 익명의 화자 벡터를 이용하여 변경된 화자 식별 벡터를 이용하여 음성 정보를 익명화할 수 있다. 이에 대해서는 하기의 설명에 의해 보다 명확하게 이해될 것이다.

[0038] 우선 이해와 설명의 편의를 도모하기 위해 CycleVAE-GAN 기반의 음성 변환 모델에 대해 간략히 설명한 후 화자 식별 벡터를 변경하는 방법에 대해 설명하기로 한다.

[0039] 본 발명의 일 실시예에 따르면, CycleVAE-GAN 기반 음성 변환 모델의 상세 구성은 도 2에 도시된 바와 같다.

[0040] 도 2를 참조하여, 음성 익명화를 위한 음성 변환 모델에 대해 보다 상세히 설명하기로 한다.

[0041] CycleVAE-GAN 기반 음성 변환 모델은 화자의 음성 정보에서 발화 정보를 유지한 상태로 익명의 화자 벡터를 이용하여 화자 정보를 제거하여 익명화된 음성 정보를 출력할 수 있다.

[0042] CycleVAE-GAN 기반 음성 변환 모델은 인코더(210), 디코더(220) 및 판별기(230)를 포함하여 구성된다. 여기서, 인코더(210) 및 디코더(220)는 VAE 기반 인코더-디코더 쌍일 수 있으며, 디코더는 GAN 모델의 생성기일 수도 있다.

[0043] 인코더(210)는 사용자의 음성 정보를 입력받아 언어 정보에 해당하는 잠재 벡터(z)를 추출하도록 훈련된다. 이때, 인코더(210)는 수학식1의 손실 함수가 최소가 되도록 훈련될 수 있다.

수학식 1

[0044]
$$\mathcal{L}_{\mathrm{VAE}}(\phi,\theta;x,X)=\mathbb{D}_{\mathrm{KL}}\left(q_{\phi}(z|x)\parallel p(z)\right)-\mathbb{E}_{z\sim q_{\phi}(z|x)}\left[\log p_{\theta}(x|z,I_X)\right]$$

[0045] 여기서, x, X, I_X, ϕ, 및 θ 는 음성 정보, 화자, 화자 식별 벡터, 인코더 파라미터, 디코더 파라미터를 나타낸다. 또한, $\mathbb{D}_{\mathrm{KL}}$ 와 $\mathbb{E}$ 는 각각 Kullback-Leibler 발산과 기대치를 나타낸다.

[0046] 인코더(210)는 수학식 1을 최소화하도록 음성 정보의 언어 정보에 해당하는 잠재 벡터(z)를 추출하도록 학습할 수 있다.

[0047] 또한, 디코더(220)는 잠재 벡터(z)와 화자 식별 벡터(Ix)를 이용하여 입력된 음성 정보를 복원하도록 훈련된다.

[0048] 본 발명의 일 실시예에 따르면, 디코더(220)는 입력된 음성 정보에서 사용자의 발화 정보(언어 정보)는 유지한 상태로 사용자의 화자 정보를 제거하기 위해 수정된 화자 식별 벡터를 이용하여 익명화할 수 있다. 화자 식별 벡터를 수정하는 방법에 대해서는 하기에서 보다 상세히 설명하기로 한다.

[0049] 순환 일관성 손실은 수학식 2와 같이 계산될 수 있다.

수학식 2

[0050]
$$\mathcal{L}_{\mathrm{Cycle}}(\phi,\theta;x,X,Y)=\mathbb{D}_{\mathrm{KL}}\left(q_{\phi}\left(z|x'_{X\to Y}\right)\parallel p(z)\right)-\mathbb{E}_{z\sim q_{\phi}(z|x'_{X\to Y})}\left[\log p_{\theta}(x|z,I_X)\right]$$

[0051] 여기서, $x'_{X\to Y}$ 는 소스 화자(X)에서 대상 화자(Y)로 변환된 음성을 나타낸다.

[0052] 화자 X의 음성 정보(x)는 대상 화자 식별 벡터 I_Y 를 사용하여 인코더와 디코더를 통과하여 x와 동일한 언어 내용을 갖지만 대상 화자(Y)의 음성으로 변환된 음성 $x'_{X\to Y}$ 이 생성할 수 있다.

[0053] 그 후 변화된 음성은 대상 화자 식별 벡터 I_Y 를 사용하여 인코더와 디코더를 거쳐 원래의 음성(x)을 복구해야 하는 변환된 역음성 $x''_{X\to Y\to X}$ 을 생성한다. 이 순환 변환은 병렬 데이터 없이 변환 경로의 명시적 학습을 장려

할 수 있다.

[0054] 화자 X와 Y에서 각각 음성 x와 y가 주어지면 CycleVAE의 손실 함수는 수학식 3과 같이 정의될 수 있다.

수학식 3

[0055]
$$\begin{aligned}\mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,y,X,Y) = {} & \mathcal{L}_{\text{VAE}}(\phi,\theta;x,X) \\ & + \mathcal{L}_{\text{VAE}}(\phi,\theta;y,Y) \\ & + \lambda_1\mathcal{L}_{\text{Cycle}}(\phi,\theta;x,X,Y) \\ & + \lambda_1\mathcal{L}_{\text{Cycle}}(\phi,\theta;y,Y,X)\end{aligned}$$

[0056] 여기서, λ_1는 CycleVAE에서 순환 일관성 손실의 가중치를 결정한다.

[0057] VAE 모듈을 최적화한 후 GAN 모듈을 사용하여 CycleVAE-GAN 기반 음성 변환 모델을 학습시킨다. VAE의 디코더는 GAN의 생성기로 간주될 수 있다. 판별기(230)는 생성기(디코더)가 대상 화자의 음성과 유사한 음성을 생성하도록 할 수 있다. 각각 화자 X 및 Y의 음성 x와 y가 주어지면, CycleVAE-GAN 기반 음성 변환 모델의 최종 손실 함수는 수학식 4와 같이 정의될 수 있다.

수학식 4

[0058]
$$\begin{aligned}&\mathcal{L}_{\text{CycleVAE-GAN}}(\phi,\theta,\psi;x,y,X,Y) \\ &= \mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,y,X,Y) \\ &\quad + \lambda_2\mathbb{E}_{y|Y}\left[D_\psi(y)\right] - \lambda_2\mathbb{E}_{z\sim q_\phi(z|x)}\left[D_\psi\left(G_\theta(z,I_Y)\right)\right] \\ &\quad + \lambda_2\mathbb{E}_{x|X}\left[D_\psi(x)\right] - \lambda_2\mathbb{E}_{z\sim q_\phi(z|y)}\left[D_\psi\left(G_\theta(z,I_X)\right)\right],\end{aligned}$$

[0059] 여기서, G_θ와 D_ψ는 각각 파라미터 θ를 가지는 생성기와 파라미터 ψ를 가지는 판별기를 나타낸다. λ_2는 CycleVAE-GAN 기반 음성 변환 모델에서 GAN 손실의 가중치를 결정한다.

[0060] 즉, 수학식 4는 VAE 및 생성기(디코더)에 대해 최소화되며, 판별기에 대해 최대화될 수 있다.

[0061] 상술한 cycleVAE-GAN 기반의 음성 변환 모델의 디코더(생성기)에 적용되는 화자 식별 벡터는 출력 음성(변환된 음성)이 대상 화자와 유사한 특성을 가지고 있음을 디코더에 안내할 수 있다.

[0062] 따라서, 화자 식별 벡터에 대해 적절한 값을 선택하면 디코더(220)가 새로운 특성을 가진 음성을 출력하도록 강제할 수 있다. 이에, 본 발명의 일 실시예에 따른 cycleVAE-GAN 기반의 음성 변환 모델의 디코더에 적용되는 화자 식별 벡터를 변경하여 사용자(즉, 대상 화자)의 언어적 내용은 유지하며 대상 화자의 신원 정보를 제거하여 낮은 화자 식별 정확도를 가지며, 높은 음성 인식 정확도를 얻을 수 있도록 할 수 있다.

[0063] 이하에서는 6개의 익명의 화자 식별 벡터에 대해 설명하기로 한다. 이해와 설명의 편의를 도모하기 위해 대상 화자 벡터와 복수의 나머지 화자(익명의 화자 벡터)의 개수가 n인 것을 가정하기로 한다. 여기서, n은 음성 변환 모델의 학습에 이용되는 화자의 수를 나타낸다.

[0064] 익명의 화자 식별 벡터는 대상 화자(사용자)와 익명의 나머지 화자 인덱스로 구성된다. 이하에서 대상 화자/대상 화자 벡터, 나머지 화자/나머지 화자 벡터는 화자 식별 벡터내의 인덱스를 지칭하는 것으로 이해되어야 할 것이다.

[0065] 익명의 화자 식별 벡터는 대상 화자와 복수의 익명의 화자를 포함하도록 구성되며, 본 발명의 일 실시예에서는 전체 화자 수가 10인 것을 가정하기로 한다. 익명의 화자의 음성 정보 또한 사전 취득되어 있는 것을 가정하기

로 한다.

[0066] 익명의 화자 식별 벡터는 6가지로 구성될 수 있다. 각각의 익명의 화자 식별 벡터는 각기 상이한 방법으로 생성될 수 있다. 이에 대해 각각 설명하기로 한다.

[0067] 제1 익명의 화자 식별 벡터

[0068] 화자 식별 벡터는 하나의 대상 화자와 9명의 익명의 화자로 구성되되, 대상 화자에 고정된 제1 값을 할당하며, 익명의 화자들은 균등 분할된 값이 할당될 수 있다. 이때, 화자 식별 벡터에 할당된 값들을 모두 합산한 결과는 "1"이 되도록 익명의 화자 벡터에 균등 분할된 값이 결정될 수 있다.

[0069] 예를 들어, 도 3의 (a)을 참조하면, 대상 화자에 고정된 "0"이 할당되는 경우, 9명의 익명의 화자에 균등 분할된 값이 할당되므로, 익명의 화자에는 "0.111"이 각각 할당될 수 있다.

[0071] 제2 익명의 화자 식별 벡터

[0072] 제1 익명의 화자 식별 벡터 생성과 유사하나 대상 화자에 할당되는 고정된 값이 상이하며, 나머지 익명의 화자에는 균등 분할된 값이 할당되는 점에서 동일하다.

[0073] 예를 들어, 도 3의 (b)에서 보여지는 바와 같이, 대상 화자에 "-1"값이 할당되며, 나머지 익명의 화자에 균등 분할된 값이 할당되되, 제2 익명의 화자 식별 벡터에 할당된 값들의 합산 결과가 1이 되어야 하므로, 익명의 화자들에 " "0.222"가 각각 할당될 수 있다.

[0075] 제3 익명의 화자 식별 벡터

[0076] 화자 식별 벡터에 포함된 익명의 화자들에 균등 분할된 값이 할당되는 점에서 제1 익명의 화자 식별 벡터 생성 방법이나 제2 화자 식별 벡터 생성 방법과 동일하다. 다만, 대상 화자에 할당되는 값이 상이하다. 즉, 대상 화자에 할당되는 값은 익명의 화자들에 균등 분할되어 할당되는 값과 절대값에서는 동일한 값이 할당되나, 익명의 화자들에 할당되는 값의 음수값이 할당된다.

[0077] 예를 들어, 대상 화자에 -A값이 할당되는 경우, 각각의 나머지 화자 에는 A값이 할당되며, 대상 화자와 각각의 나머지 화자에 할당된 값들을 합산한 결과가 1이여야 하므로, 나머지 화자에는 각각 0.125값이 할당되며, 대상 화자는 각 나머지 화자 벡터에 할당된 값의 음수값인 -0.125값이 할당될 수 있다.

[0078] 이와 같이, 제1 익명의 화자 식별 벡터 내지 제3 익명의 화자 식별 벡터는 화자 식별 벡터에 포함되는 복수의 익명의 화자에 각각 균등 분할된 값이 할당되는 점에서 동일하나 대상 화자에 할당되는 값에 따라 익명의 화자에 균등 분할되는 값이 상이하게 결정될 수 있다.

[0079] 제4 화자 익명의 식별 벡터 내지 제6 익명의 화자 식별 벡터는 코사인 유사도를 이용한다.

[0080] 즉, 대상 화자의 음성 정보와 익명의 화자의 음성 정보에서 i-벡터를 각각 추출하며, i-벡터를 이용하여 대상 화자와 나머지 익명의 화자간의 코사인 유사도를 도출할 수 있다. 제4 익명의 화자 식별 벡터 내지 제6 익명의 화자 식별 벡터는 이와 같이 도출된 코사인 유사도를 이용하여 화자 벡터에 각각 상이한 값을 할당할 수 있다.

[0082] 제4 익명의 화자 식별 벡터

[0083] 대상 화자에는 고정된 값(예를 들어, 0)이 할당되며, 익명의 화자는 각각 코사인 유사도의 역에 따른 상대적 거리에 따라 0~1 사이에서 정규화된 값이 각각 할당되어 화자 식별 벡터가 생성(수정) 될 수 있다(도 3의 (d) 참조). 화자 식별 벡터에 할당된 값들을 합산한 결과는 "1"이 되도록 익명의 화자에 할당되는 값들이 결정될 수 있다.

[0085] 제5 익명의 화자 식별 벡터

[0086] 대상 화자에 고정된 값(예를 들어, 0)이 할당되며, 익명의 화자 중 코사인 유사도가 가장 먼 화자에만 고정된 "1"을 할당하여 화자 식별 벡터가 생성(변경)될 수 있다(도 3의 (e) 참조).

[0087] 제6 익명의 화자 식별 벡터

[0088] 대상 화자에 고정된 값(예를 들어, 0)이 할당되며, 익명의 화자 중 코사인 유사도가 가장 먼 순으로 두명의 화자에만 고정된 값(0.5) 을 할당하여 화자 식별 벡터가 생성될 수 있다(도 3의 (f) 참조).

[0089] 익명화부(120)는 제1 내지 제6 익명의 화자 식별 벡터를 이용하여 음성 변환 모델의 디코더단에 각각 적용하여 사용자의 음성 정보에서 발화 정보를 유지한 상태로 화자 정보만 익명화한 변환된 음성 정보를 출력할 수 있다.

[0090] 입력부(125)는 사용자로부터 음성 익명화 장치(100)의 제어를 위한 명령을 입력받기 위한 수단이다.

[0091] 음성 출력부(130)는 음성을 출력하기 위한 수단이다. 음성 출력부(130)는 예를 들어, 스피커일 수 있다. 즉, 음성 출력부(130)는 프로세서(140)의 제어에 따라 제1 내지 제6 익명의 화자 식별 벡터를 이용하여 익명화된 음성 정보를 출력할 수 있다.

[0092] 이에 따라, 사용자는 제1 내지 제6 익명의 화자 식별 벡터에 의해 각각 익명화된 음성 정보를 직접 들은 후 자신의 신원 정보가 가장 제거가 잘된 음성을 입력부(125)를 통해 선택하여 서버로 전송할 수도 있다.

[0093] 메모리(135)는 본 발명의 일 실시예에 따른 익명의 화자 벡터를 이용하여 사용자의 음성 정보를 익명화하는 방법을 수행하기 위한 다양한 명령어들을 저장하기 위한 수단이다.

[0094] 프로세서(140)는 본 발명의 일 실시예에 따른 음성 익명화 장치(100)의 내부 구성 요소들(예를 들어, 통신부(110), 음성 입력부(115), 익명화부(120), 입력부(125), 음성 출력부(130), 메모리(135) 등)을 제어하기 위한 수단이다.

[0095] 또한, 프로세서(140)는 제1 내지 제6 익명의 화자 식별 벡터를 이용하여 각각 익명화된 음성 중 사용자에 의해 선택된 음성을 서버로 전송하도록 통신부(110)를 제어할 수도 잇다.

[0097] 도 4는 본 발명의 일 실시예에 따른 음성 익명화 방법을 나타낸 순서도이다.

[0098] 단계 410에서 음성 익명화 장치(100)는 음성 정보를 취득한다. 즉, 마이크를 통해 사용자의 음성 정보를 취득할 수 있다.

[0099] 단계 415에서 음성 익명화 장치(100)는 취득한 음성 정보를 음성 변환 모델에 적용하여 변환한다. 이때, 음성 익명화 장치(100)는 익명의 화자 식별 벡터(제1 익명의 화자 식별 벡터 내지 제6 익명의 화자 식별 벡터)를 반영하여 사용자의 음성 정보에서 발화 정보를 유지한 상태로 화자 정보를 제거하여 익명화할 수 있다. 이에 대해서는 도 1 내지 도 3을 참조하여 설명한 바와 동일하므로 중복되는 설명은 생략하기로 한다. 단계 420에서 음성 익명화 장치(100)는 익명화된 음성 정보 중 어느 하나를 선택받는다.

[0100] 단계 425에서 음성 익명화 장치(100)는 사용자에 의해 선택된 음성 정보(익명화된 음성 정보)를 서버로 전송한다.

[0101] 본 발명의 일 실시예에 따르면, 음성 익명화 장치(100)는 제1 내지 제6 익명의 화자 식별 벡터를 적용하여 익명화된 음성 정보를 각각 사용자에게 제공한 후 어느 하나를 선택받을 수 있다. 즉, 사용자는 제1 내지 제6 익명의 화자 식별 벡터를 이용하여 변환된 음성 정보(익명화된 음성 정보)를 직접 청취한 후 가장 잘 익명화된 음성 정보를 서버로 전송하도록 할 수 있는 이점이 있다. 또한, 본 발명은 음성 변환 모델의 디코더에서 제1 내지 제6 익명의 화자 식별 벡터를 이용하여 음성 변환에 이용함으로써 소리 신호 수준에서 익명화를 적용할 수 있는 이점도 있다.

[0102] 도 5는 익명의 화자 식별 벡터에 대한 화자 식별 및 음성 인식 정확도를 나타낸 결과이다.

[0103] 도 5에서 보여지는 바와 같이, 제1 내지 제6 익명의 화자 식별 벡터가 사용자의 신원을 성공적으로 제거한 것을 알 수 있다. 또한, 자체 변환된 음성 데이터에 대한 음성 인식 정확도가 73.4%라는 점을 고려하면, 본 발명의 일 실시예에 따른 6가지 화자 식별 벡터에 의한 음성 변환시 언어 콘텐츠를 더 이상 손상시키지 않는 것을 알 수 있다.

[0105] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데

이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0106] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0107] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0109] 100: 음성 익명화 장치

110: 통신부

115: 음성 입력부

120: 익명화부

125: 입력부

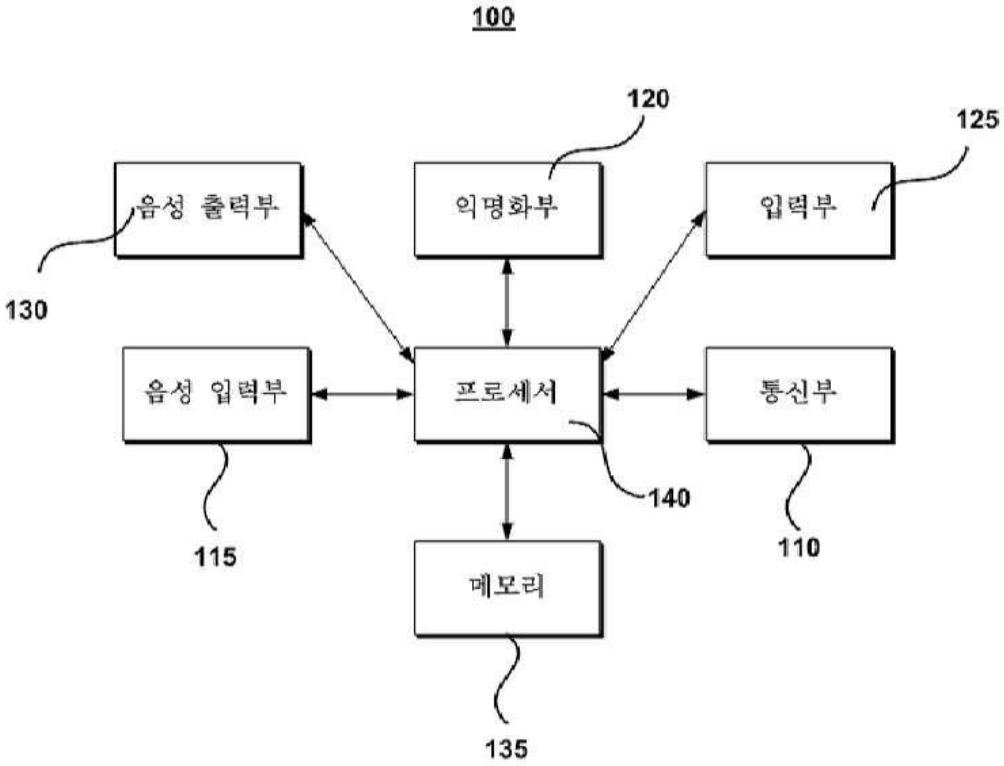
130: 출력부

135: 메모리

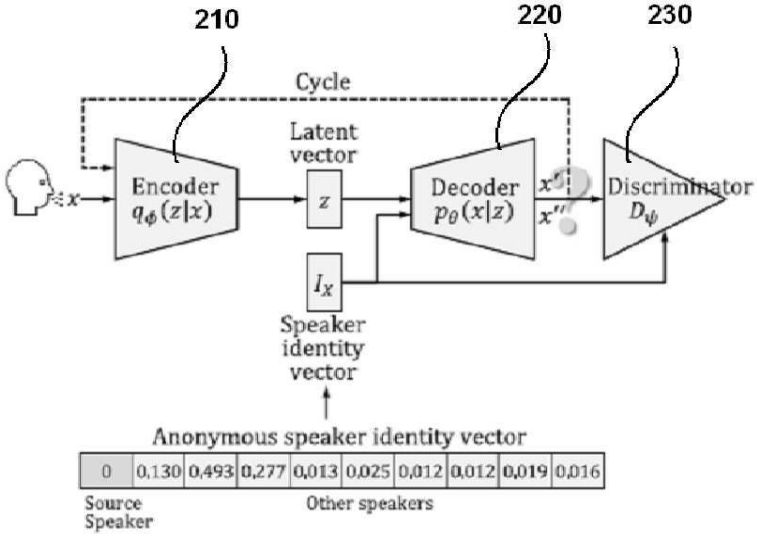
140: 프로세서

도면

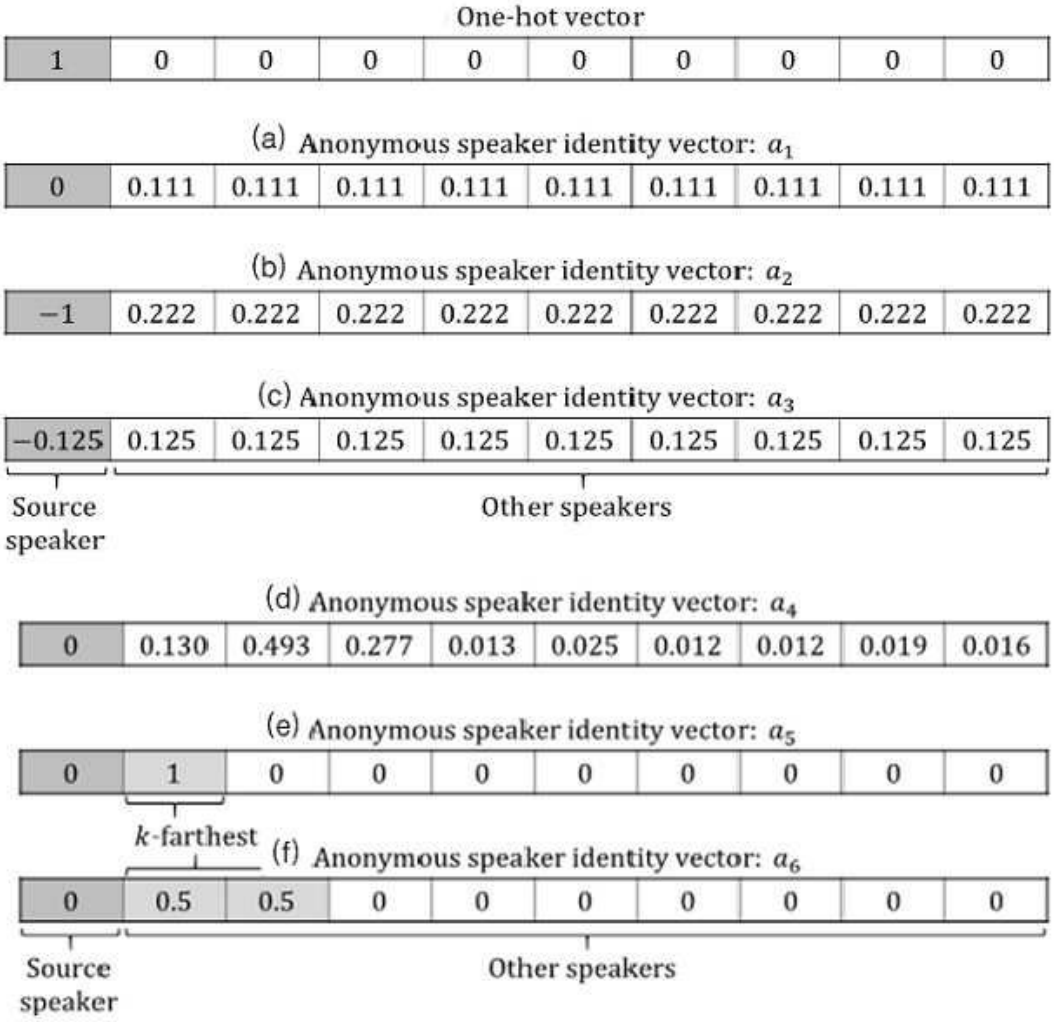
도면1


100
120
125
음성 출력부
익명화부
입력부
130
음성 입력부
프로세서
통신부
140
115
110
메모리
135

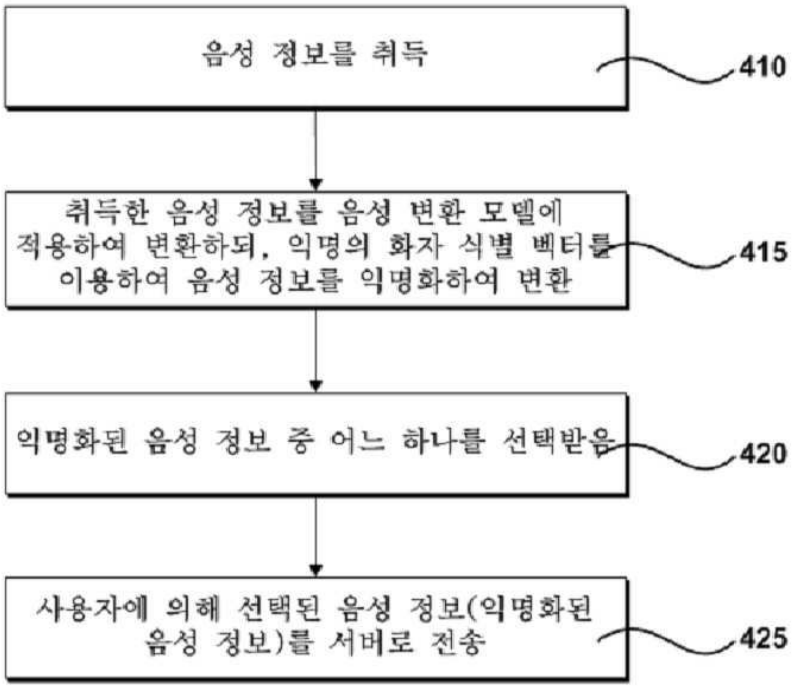
도면2


210
220
230
Cycle
Latent vector
Encoder $q_\phi(z|x)$
z
Decoder $p_\theta(x|z)$
x'
x''
Discriminator D_ψ
x
I_x
Speaker identity vector
Anonymous speaker identity vector
0
0.130
0.493
0.277
0.013
0.025
0.012
0.012
0.019
0.016
Source Speaker
Other speakers

도면3


One-hot vector
1 0 0 0 0 0 0 0 0 0
(a) Anonymous speaker identity vector: a_1
0 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111
(b) Anonymous speaker identity vector: a_2
−1 0.222 0.222 0.222 0.222 0.222 0.222 0.222 0.222 0.222
(c) Anonymous speaker identity vector: a_3
−0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125
Source speaker
Other speakers
(d) Anonymous speaker identity vector: a_4
0 0.130 0.493 0.277 0.013 0.025 0.012 0.012 0.019 0.016
(e) Anonymous speaker identity vector: a_5
0 1 0 0 0 0 0 0 0 0
k-farthest
(f) Anonymous speaker identity vector: a_6
0 0.5 0.5 0 0 0 0 0 0 0
Source speaker
Other speakers

도면4



도면5

Speaker identity vector	Source speaker value	Non-source speaker value	Speaker identification accuracy (%)		Speech recognition accuracy (%)
			GMM	DNN	
a_1	0	$1/(n-1)$	18.8	7.0	69.3
a_2	-1	$2/(n-1)$	0.7	**0.1**	57.0
a_3	$-1/(n-2)$	$1/(n-2)$	11.8	5.7	68.5
a_4	0	Inverse of cosine similarity	15.9	9.2	**72.1**
a_5	0	1 for the 1-farthest speaker and 0 for others	**0.3**	**0.1**	69.3
a_6	0	0.5 for the 2-farthest speakers and 0 for others	20.9	9.6	69.2