

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0030358
(43) 공개일자 2025년03월05일

(51) 국제특허분류(Int. Cl.)

G10L 13/08 (2006.01) G06N 3/0464 (2023.01)
G10L 13/02 (2006.01) G10L 15/02 (2006.01)
G10L 15/14 (2006.01) G10L 15/26 (2006.01)
G10L 17/02 (2013.01) G10L 25/18 (2013.01)
G10L 25/30 (2013.01) G10L 25/51 (2013.01)

(52) CPC특허분류

G10L 13/08 (2013.01)
G06N 3/0464 (2023.01)

(21) 출원번호 10-2023-0134905

(22) 출원일자 2023년10월11일

심사청구일자 2023년10월11일

(30) 우선권주장

1020230111192 2023년08월24일 대한민국(KR)

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

이성환

서울특별시 강남구 언주로30길 13, B-2506(도곡동, 대림아크로빌, 대림아크로텔)

배현우

대구광역시 북구 호암로 20, 102동 1605호(칠성동2가 성광우방타운)

이상훈

서울특별시 성북구 인촌로27길 34-8 (안암동5가)

(74) 대리인

윤귀상

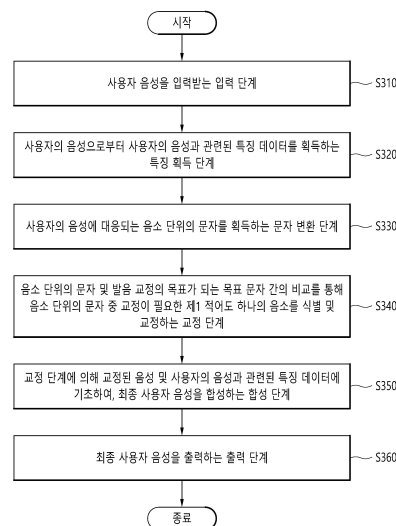
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 영어 원어된 수준의 발음 교정이 가능한 음성 변환 동작을 수행하는 장치 및 방법

(57) 요약

사용자의 입력 발음의 교정을 위한 발음 변환 방법 및 장치가 개시된다. 본 개시의 일 실시예에 따른, 장치에 의해 수행되는, 사용자의 입력 발음의 교정을 위한 발음 변환 방법은, 사용자 음성을 입력받는 입력 단계; 상기 사용자의 음성으로부터 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 특징 획득 단계; 상기 사용자의 음성으로 대응되는 음소 단위의 문자를 획득하는 문자 변환 단계; 상기 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교를 통해 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 식별 및 교정하는 교정 단계; 상기 교정 단계에 의해 교정된 음성 및 상기 사용자의 음성과 관련된 특징 데이터에 기초하여, 최종 사용자 음성을 합성하는 합성 단계; 및 상기 최종 사용자 음성을 출력하는 출력 단계를 포함할 수 있다.

대표도 - 도3



(52) CPC특허분류

G10L 13/02 (2013.01)
 G10L 15/14 (2013.01)
 G10L 15/26 (2013.01)
 G10L 17/02 (2013.01)
 G10L 25/18 (2013.01)
 G10L 25/30 (2013.01)
 G10L 25/51 (2013.01)
 G10L 2015/025 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193973
과제번호	2019-0-00079-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(고려대학교)
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2019.04.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193622
과제번호	2021-0-02068-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능 혁신 허브 연구 개발
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2021.07.01 ~ 2023.12.31

명세서

청구범위

청구항 1

장치에 의해 수행되는, 사용자의 입력 발음의 교정을 위한 발음 변환 방법에 있어서, 상기 방법은:

사용자 음성을 입력받는 입력 단계;

상기 사용자의 음성으로부터 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 특징 획득 단계;

상기 사용자의 음성에 대응되는 음소 단위의 문자를 획득하는 문자 변환 단계;

상기 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교를 통해 상기 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 식별 및 교정하는 교정 단계;

상기 교정 단계에 의해 교정된 음성 및 상기 사용자의 음성과 관련된 특징 데이터에 기초하여, 최종 사용자 음성을 합성하는 합성 단계; 및

상기 최종 사용자 음성을 출력하는 출력 단계를 포함하는, 방법.

청구항 2

제1항에 있어서,

상기 교정 단계는,

상기 음소 단위의 문자를 편차 모델에 입력하여 상기 음소 단위의 문자 중 교정이 필요한 제2 적어도 하나의 음소를 식별 및 교정하는 단계를 포함하는, 방법.

청구항 3

제2항에 있어서,

상기 편차 모델은, 복수의 음소 유형에 대한 정답 발음 및 상기 복수의 음소 유형에 대해 복수의 사용자의 실제 발음에 기초하여, 상기 음소 단위의 문자 중 상기 복수의 음소 유형에 대응되는 상기 제2 적어도 하나의 음소를 식별하도록 학습되는, 방법.

청구항 4

제3항에 있어서,

상기 사용자의 음성과 관련된 특징 데이터, 상기 교정된 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 음성 합성 모델에 입력하여 상기 최종 사용자 음성이 출력되는, 방법.

청구항 5

제4항에 있어서,

상기 출력 단계는,

상기 음소 단위의 문자로부터 교정된 상기 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 나타내는 정보를 표시하는, 방법.

청구항 6

제1항에 있어서,

상기 특징 획득 단계는,

상기 사용자의 음성에 대한 멜-스펙트로그램(mel-spectrogram)을 CNN(convolution neural network) 모델 또는 어텐션 알고리즘 기반의 모델 중의 적어도 하나에 입력하여 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 단계를 포함하는, 방법.

청구항 7

사용자의 입력 발음의 교정을 위한 발음 변환하는 장치에 있어서, 상기 장치는:

사용자 음성을 입력받는 음성 입력부;

상기 사용자의 음성으로부터 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 화자 특징 추출부;

상기 사용자의 음성에 대응되는 음소 단위의 문자를 획득하는 음성 인식부;

상기 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교를 통해 상기 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 식별 및 교정하는 발음 교정부;

상기 발음 교정부에 의해 교정된 음성 및 상기 사용자의 음성과 관련된 특징 데이터에 기초하여, 최종 사용자 음성을 합성하는 음성 합성부; 및

상기 최종 사용자 음성을 출력하는 음성 출력부를 포함하는, 장치.

청구항 8

제7항에 있어서,

상기 발음 교정부는,

상기 음소 단위의 문자를 편차 모델에 입력하여 상기 음소 단위의 문자 중 교정이 필요한 제2 적어도 하나의 음소를 식별 및 교정하는, 장치.

청구항 9

제8항에 있어서,

상기 편차 모델은, 복수의 음소 유형에 대한 정답 발음 및 상기 복수의 음소 유형에 대해 복수의 사용자의 실제 발음에 기초하여, 상기 음소 단위의 문자 중 상기 복수의 음소 유형에 대응되는 상기 제2 적어도 하나의 음소를 식별하도록 학습되는, 장치.

청구항 10

제9항에 있어서,

상기 사용자의 음성과 관련된 특징 데이터, 상기 교정된 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 음성 합성 모델에 입력하여 상기 최종 사용자 음성이 출력되는, 장치.

청구항 11

제10항에 있어서,

상기 음성 출력부는,

상기 음소 단위의 문자로부터 교정된 상기 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 나타내는 정보를 표시하는, 장치.

청구항 12

제1항에 있어서,

상기 화자 특징 추출부는,

상기 사용자의 음성에 대한 멜-스펙트로그램(mel-spectrogram)을 CNN(convolution neural network) 모델 또는 어텐션 알고리즘 기반의 모델 중의 적어도 하나에 입력하여 상기 사용자의 음성과 관련된 특징 데이터를 획득하는, 장치.

발명의 설명

기술 분야

[0001] 본 개시는 영어 원어민 수준의 발음 교정이 가능한 음성 변환 기술에 관한 것으로서, 보다 상세하게는 음성 인식 모델을 활용하여 입력되는 사용자의 발음을 인식하고, 잘못 발음된 음소에 대응되는 음성 및 영어 원어민 수준의 정확한 음소의 발음을 합성하는 방법 및 장치에 관한 것이다.

배경 기술

[0002] 영어 교육 활용을 위해 사용자의 목소리는 유지하고 발음과 관련된 정보는 교정하는 기술이 필요하다. 종래에는 사용자의 발음 교정을 위하여 음성 인식(automatic recognition, ASR), 음성 변환(voice conversion, VC), 말하는 스타일의 변환(speaking style conversion, SSC) 기술 등이 활용되었다.

[0003] ASR 기술은 사용자의 발화를 인식하고 이를 원하는 문자의 형태로 시각화하기 위해 사용되는 기술이다. 사용자의 발화를 정확하게 인식하여 이를 사용자가 원하는 방향으로 문자 변환을 수행하는 것이 중요하다. 현재 이용되고 있는 ASR 모델은 단어 단위의 인식 결과를 출력하거나, 문자 단위의 인식 결과를 보여주는 경우가 대부분이다.

[0004] VC 기술은 입력 음성에서 언어적 정보는 유지한 채, 원하는 목소리와 스타일로 변환하는 기술이다. 사용자의 발화 중 변환이 필요한 정보와 변환이 필요하지 않은 언어적 정보에 대해 정확히 분리하는 것이 중요하다. 여기서, 정보를 분리하는 기술은 발음 교정 기술에 이용될 수 있다.

[0005] SSC 기술은 입력 음성에서 언어적 정보는 유지한 채 원하는 목소리와 스타일로 변환하는 기술이다. SSC 기술은 VC 기술에 더해 발화의 길이나 억양과 같은 스타일과 관련된 정보를 고려할 수 있으며, 사용자가 원하는 길이나 억양을 가진 스타일로의 변환이 가능하게 하는 기술이다.

[0006] 한편, 종래에는 발화된 음성에 대해 문맥을 파악하거나 연속된 단어들에 대한 인식을 통해 음성을 정확하게 인식하고자 하는 것에 초점을 맞췄다. 다만, 종래 기술들은 단어나 문자 단위에서 잘못된 발음에 대해 인식할 때 인식된 단어나 문자가 실제로 존재하는 경우, 해당 발음이 실제 정확한 발음인지 아닌지를 명확히 판단하지 못한다는 문제점이 존재하였다.

발명의 내용

해결하려는 과제

[0007] 본 개시의 목적은 발화자의 목소리는 유지하되 발음과 관련된 정보는 교정하는 방법 및 장치가 개시된다.

[0008] 본 개시의 목적은 음소 단위의 인식 모델을 이용하여 정확한 발음을 인식하여 발화자의 목소리 특징을 추출하고, 인식된 음소와 발화자의 목소리 특징을 이용하여 발화자의 음성을 합성하는 방법 및 장치가

개시된다.

- [0009] 본 개시의 목적은 발화자가 자주 틀리는 발음들에 대해 편차 모델을 이용하여 더 정확한 발음의 음성을 합성하는 방법 및 장치가 개시된다.
- [0010] 본 개시의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 개시의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 개시의 목적 및 장점들은 청구범위에 나타난 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

- [0011] 본 개시의 일 실시예에 따른, 장치에 의해 수행되는, 사용자의 입력 발음의 교정을 위한 발음 변환 방법은, 사용자 음성을 입력받는 입력 단계; 상기 사용자의 음성으로부터 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 특징 획득 단계; 상기 사용자의 음성에 대응되는 음소 단위의 문자를 획득하는 문자 변환 단계; 상기 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교를 통해 상기 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 식별 및 교정하는 교정 단계; 상기 교정 단계에 의해 교정된 음성 및 상기 사용자의 음성과 관련된 특징 데이터에 기초하여, 최종 사용자 음성을 합성하는 합성 단계; 및 상기 최종 사용자 음성을 출력하는 출력 단계를 포함할 수 있다.
- [0012] 그리고, 상기 교정 단계는, 상기 음소 단위의 문자를 편차 모델에 입력하여 상기 음소 단위의 문자 중 교정이 필요한 제2 적어도 하나의 음소를 식별 및 교정하는 단계를 포함할 수 있다.
- [0013] 그리고, 상기 편차 모델은, 복수의 음소 유형에 대한 정답 발음 및 상기 복수의 음소 유형에 대해 복수의 사용자의 실제 발음에 기초하여, 상기 음소 단위의 문자 중 상기 복수의 음소 유형에 대응되는 상기 제2 적어도 하나의 음소를 식별하도록 학습될 수 있다.
- [0014] 그리고, 상기 사용자의 음성과 관련된 특징 데이터, 상기 교정된 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 음성 합성 모델에 입력하여 상기 최종 사용자 음성이 출력될 수 있다.
- [0015] 그리고, 상기 출력 단계는, 상기 음소 단위의 문자로부터 교정된 상기 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 나타내는 정보를 표시할 수 있다.
- [0016] 그리고, 상기 특징 획득 단계는, 상기 사용자의 음성에 대한 멜-스펙트로그램(mel-spectrogram)을 CNN(convolution neural network) 모델 또는 어텐션 알고리즘 기반의 모델 중의 적어도 하나에 입력하여 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 단계를 포함할 수 있다.
- [0017] 본 개시의 일 실시예로, 사용자의 입력 발음의 교정을 위한 발음 변환하는 장치는, 사용자 음성을 입력받는 음성 입력부; 상기 사용자의 음성으로부터 상기 사용자의 음성과 관련된 특징 데이터를 획득하는 화자 특징 추출부; 상기 사용자의 음성에 대응되는 음소 단위의 문자를 획득하는 음성 인식부; 상기 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교를 통해 상기 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 식별 및 교정하는 발음 교정부; 상기 발음 교정부에 의해 교정된 음성 및 상기 사용자의 음성과 관련된 특징 데이터에 기초하여, 최종 사용자 음성을 합성하는 음성 합성부; 및 상기 최종 사용자 음성을 출력하는 음성 출력부를 포함할 수 있다.
- [0018] 그리고, 상기 발음 교정부는, 상기 음소 단위의 문자를 편차 모델에 입력하여 상기 음소 단위의 문자 중 교정이 필요한 제2 적어도 하나의 음소를 식별 및 교정할 수 있다.
- [0019] 그리고, 상기 음성 출력부는, 상기 음소 단위의 문자로부터 교정된 상기 제1 적어도 하나의 음소의 발음 및 상기 제2 적어도 하나의 음소의 발음을 나타내는 정보를 표시할 수 있다.
- [0020] 그리고, 상기 화자 특징 추출부는, 상기 사용자의 음성에 대한 멜-스펙트로그램(mel-spectrogram)을 CNN(convolution neural network) 모델 또는 어텐션 알고리즘 기반의 모델 중의 적어도 하나에 입력하여 상기 사용자의 음성과 관련된 특징 데이터를 획득할 수 있다.
- [0021] 상기 과제의 해결 수단은, 본 개시의 특징을 모두 열거한 것은 아니다. 본 개시의 다양한 특징과 그에 따른 장점과 효과는 아래의 구체적인 실시예를 참조하여 보다 상세하게 이해될 수 있을 것이다.

발명의 효과

[0022] 본 개시의 다양한 실시예에 의해, 발화자는 영어 원어민 수준의 발음 교정이 가능한 음성 변환 기술을 제공받을 수 있다.

[0023] 본 개시의 다양한 실시예에 의해, 발화자는 보다 자연스럽게 정확한 발음으로 변환된 음성을 제공받을 수 있다.

[0024] 본 개시의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0025] 도 1은 본 개시의 일 실시예에 따른, 영어 원어민 수준의 발음 교정이 가능한 음성 변환 동작을 수행하는 장치의 구성을 설명하기 위한 블록도이다.

도 2는 본 개시의 일 실시예에 따른, 발음 교정부의 구성 및 동작을 설명하기 위한 도면이다.

도 3은 본 개시의 일 실시예에 따른, 영어 원어민 수준의 발음 교정이 가능한 음성 변환 동작을 수행하는 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0026] 본 개시의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시의 개시가 완전하도록 하고, 본 개시가 속하는 기술 분야의 통상의 기술자에게 본 개시의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 개시는 청구항의 범주에 의해 정의될 뿐이다.

[0027] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 개시를 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다.

[0028] 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 개시의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0029] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 개시가 속하는 기술 분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0030] 공간적으로 상대적인 용어인 "아래(below)", "아래(beneath)", "하부(lower)", "위(above)", "상부(upper)" 등은 도면에 도시되어 있는 바와 같이 하나의 구성요소와 다른 구성요소와의 상관관계를 용이하게 기술하기 위해 사용될 수 있다. 공간적으로 상대적인 용어는 도면에 도시되어 있는 방향에 더하여 사용시 또는 동작 시 구성요소들의 서로 다른 방향을 포함하는 용어로 이해되어야 한다.

[0031] 예를 들어, 도면에 도시되어 있는 구성요소를 뒤집을 경우, 다른 구성요소의 "아래(below)"또는 "아래(beneath)"로 기술된 구성요소는 다른 구성요소의 "위(above)"에 놓여질 수 있다. 따라서, 예시적인 용어인 "아래"는 아래와 위의 방향을 모두 포함할 수 있다. 구성요소는 다른 방향으로도 배향될 수 있으며, 이에 따라 공간적으로 상대적인 용어들은 배향에 따라 해석될 수 있다.

[0032] 이하에서는 도면을 참조하여 영어 원어민 수준의 발음 교정이 가능한 음성 변환 동작을 수행하는 장치 및 방법을 설명하도록 한다.

[0034] 도 1은 본 개시의 일 실시예에 따른, 영어 원어민 수준의 발음 교정이 가능한 음성 변환 동작을 수행하는 장치의 구성을 설명하기 위한 블록도이다.

- [0035] 도 1에서 설명하는 장치는 스마트 폰, 태블릿 PC, 웨어러블 장치, 노트북 등을 포함할 수 있으나 이에 제한되는 것은 아니다.
- [0036] 도 1에 도시된 바와 같이, 장치(100)는 메모리(110), 통신 모듈(120), 마이크(130) 및 프로세서(140)를 포함할 수 있다. 다만, 이에 국한되는 것은 아니며, 장치(100)는 필요한 동작에 따라 당업자 관점에서 자명한 범위 내에서 소프트웨어 및 하드웨어 구성이 수정/추가/생략될 수 있다.
- [0037] 메모리(110)는 본 장치(100)의 다양한 기능을 지원하는 데이터와, 제어부의 동작을 위한 프로그램을 저장할 수 있고, 입/출력되는 데이터들을 저장할 있고, 본 장치에서 구동되는 다수의 응용 프로그램(application program 또는 애플리케이션(application)), 본 장치의 동작을 위한 데이터들, 명령어들을 저장할 수 있다. 이러한 응용 프로그램 중 적어도 일부는, 무선 통신을 통해 외부 장치로부터 다운로드 될 수 있다.
- [0038] 이러한, 메모리(110)는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), SSD 타입(Solid State Disk type), SDD 타입(Silicon Disk Drive type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(random access memory; RAM), SRAM(static random access memory), 롬(read-only memory; ROM), EEPROM(electrically erasable programmable read-only memory), PROM(programmable read-only memory), 자기 메모리, 자기 디스크 및 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 또한, 메모리(110)는 본 장치와는 분리되어 있으나, 유선 또는 무선으로 연결된 데이터베이스가 될 수도 있다.
- [0039] 통신 모듈(120)은 외부 장치와 통신을 가능하게 하는 하나 이상의 구성 요소를 포함할 수 있다. 예를 들어, 통신 모듈(120)은 무선통신 모듈, 유선통신 모듈 또는 위치정보 모듈 중 적어도 하나를 포함할 수 있다.
- [0040] 여기서, 무선 통신 모듈은 와이파이(Wifi) 모듈, 와이브로(Wireless broadband) 모듈 외에도, GSM(global System for Mobile Communication), CDMA(Code Division Multiple Access), WCDMA(Wideband Code Division Multiple Access), UMTS(universal mobile telecommunications system), TDMA(Time Division Multiple Access), LTE(Long Term Evolution), 4G, 5G, 6G 등 다양한 무선 통신 방식을 지원하는 무선 통신 모듈을 포함할 수 있다.
- [0041] 마이크(130)는 사용자(즉, 발화자)의 음성을 입력받을 수 있다. 도 1에서는 마이크(130)가 장치(100)의 일 구성 요소로 구현된 경우를 도시하고 있으나 이에 제한되는 것은 아니다. 마이크(130)는 장치(100)의 외부에 구성될 수 있으며, 장치(100)는 마이크(130)로부터 입력된 발화자의 음성을 통신 모듈(120)을 통해 입력받을 수 있다.
- [0042] 프로세서(140)는 본 장치(100) 내의 구성요소들의 동작을 제어하기 위한 알고리즘 또는 알고리즘을 재현한 프로그램에 대한 데이터를 저장하는 메모리, 및 메모리에 저장된 데이터를 이용하여 전술한 동작을 수행하는 적어도 하나의 프로세서(미도시)로 구현될 수 있다. 이때, 메모리와 프로세서는 각각 별개의 칩으로 구현될 수 있다. 또는, 메모리와 프로세서는 단일 칩으로 구현될 수도 있다.
- [0043] 즉, 프로세서(140)는 본 장치(100) 내의 구성 요소들을 제어함으로써 장치의 전반적인 동작 및 기능을 제어할 수 있다. 프로세서는 이하의 도면에서 설명되는 본 개시에 따른 다양한 실시 예들을 본 장치 상에서 구현하기 위하여, 위에서 살펴본 구성요소들을 중 어느 하나 또는 복수를 조합하여 제어할 수 있다.
- [0044] 본 개시의 일 예로, 도 1에 도시된 바와 같이, 프로세서(140)는 음성 입력부(140-1), 화자 특징 추출부(140-2), 음성 인식부(140-3), 발음 교정부(140-4), 음성 합성부(140-5) 및 음성 출력부(140-6)를 포함할 수 있다. 즉, 프로세서(140)는 음성 입력부(140-1), 화자 특징 추출부(140-2), 음성 인식부(140-3), 발음 교정부(140-4), 음성 합성부(140-5) 및 음성 출력부(140-6)를 제어할 수 있다.
- [0045] 음성 입력부(140-1)는 마이크(130)를 통해 사용자의 음성을 입력받을 수 있다. 음성 입력부(140-1)는 입력된 사용자의 음성을 음성 파일 형태로 메모리(110)에 저장할 수 있다.
- [0046] 화자 특징 추출부(140-2)는 음성 입력부(140-1)를 통해 입력된 사용자의 음성에 대한 특징 데이터(즉, 사용자의 목소리 특징 데이터)를 추출할 수 있다. 사용자의 음성에 대한 특징 데이터는 사용자의 음성의 스타일 정보를 포함할 수 있다.
- [0047] 일 예로, 화자 특징 추출부(140-2)는 사용자의 음성으로부터 사용자의 음성에 대한 특징 데이터를 추출할 수 있는 심층 학습 계층 구조를 가질 수 있다.
- [0048] 구체적으로, 화자 특징 추출부(140-2)는 사용자의 음성으로부터 변환된 멜-스펙트로그램(mel-spectrogram)로부

터 사용자의 음성에 대한 특징 데이터를 추출할 수 있다. 이 때, 화자 특징 추출부(140-2)는 CNN(convolution) 모델 및 어텐션(attention) 알고리즘 기반의 모델 중의 적어도 하나를 이용하여 사용자의 음성으로부터 특징 데이터를 추출할 수 있다.

- [0049] 음성 인식부(140-3)는 입력된 음성에 대해 음소 단위의 문자를 출력할 수 있다. 구체적으로, 음성 인식부(140-3)는 음소 단위의 인식 모델을 이용하여 입력된 음성에 대응되는 음소 단위의 문자를 출력할 수 있다.
- [0050] 음성 인식부(140-3)로부터 출력된 음소 단위의 문자는 발음 교정부(140-4)에 입력되어 교정 작업이 진행될 수 있다. 즉, 발음 교정부(140-4)는 음성 인식부(140-3)를 통해 인식된 문자에 대해 교정된 발음의 문자로 교정할 수 있다. 발음 교정부(140-4)의 구성 및 동작은 도 2를 참조하여 구체적으로 설명하도록 한다.
- [0051] 음성 합성부(140-5)는 발음 교정부(140-4)에 의해 교정된 발음 결과 및 화자 특징 추출부(140-2)를 통해 추출된 사용자의 음성에 대한 특징 데이터에 기초하여 음성을 합성할 수 있다.
- [0052] 구체적으로, 음성 합성부(140-5)는 발음 교정부(140-4)에 의해 교정된 발음 결과 및 화자 특징 추출부(140-2)를 통해 추출된 사용자의 음성에 대한 특징 데이터를 음성 합성 모델에 입력하여 합성된 음성을 출력할 수 있다.
- [0053] 음성 출력부(140-6)는 음성 합성부(140-5)를 통해 획득된 음성을 출력할 수 있다. 여기서, 음성 출력부(140-6)는 음성 합성부(140-5)를 통해 획득된 음성에 대한 음성 파일을 생성하고 이를 메모리(110)에 저장할 수 있다.
- [0055] 도 2는 본 개시의 일 실시예에 따른, 발음 교정부의 구성 및 동작을 설명하기 위한 도면이다.
- [0056] 도 2에 도시된 바와 같이, 발음 교정부(140-4)는 문자 입력부(141), 목표 문자 입력부(142), 편차 모델(143), 문자 출력부(144) 및 교정 확인부(145)를 포함할 수 있다.
- [0057] 문자 입력부(141)는 음성 인식부(140-3)를 통해 인식된 문자를 입력받을 수 있다. 목표 문자 입력부(142)는 음소 단위로 변환된 발음 교정의 목표가 되는 문자를 입력받을 수 있다.
- [0058] 발음 교정부(140-4)는 음성 인식부(140-3)를 통해 입력된 문자와 목표 문자 입력부(142)를 통해 입력된 발음 교정의 목표가 되는 문자를 비교함으로써 교정이 필요한 음소 부분 식별 및 교정할 수 있다.
- [0059] 이 때, 발음 교정부(140-4)는 편차 모델(143)을 이용하여 일반적으로 자주 틀리는 발음에 대한 정확한 교정을 음성 인식부(140-3)를 통해 인식된 문자에 대해 수행할 수 있다.
- [0060] 여기서, 편차 모델(143)은 일반적으로 자주 틀리거나 헛갈리는 발음에 대해 문맥에 기초하여 정확한 발음을 제시할 수 있다. 편차 모델(143)은 L2-Artic 데이터 셋을 활용하여 학습될 수 있다.
- [0061] L2-Artic 데이터 셋은 비영어권 성인 영어 발음에 기초한 데이터 셋이다. 구체적으로, L2-Artic 데이터 셋은 목표 문자 기반의 정확한 발음을 나타낸 음소 문자 및 실제로 발화자의 발음을 나타내는 음소 문자를 포함할 수 있다.
- [0062] 예로, 영어 음소의 개수는 43~45 개로 표현될 수 있다. L2-Artic 데이터 셋으로 학습된 CNN을 포함하는 편차 모델(143)은 특정 입력 음소에 대해서 화자들이 자주 틀리는 음소에 대해 정확한 발음의 음소를 제공할 수 있도록 도와주는 역할을 담당할 수 있다.
- [0063] 일 예로, 편차 모델(143)은 음성 인식부(140-3)를 통해 인식된 문자 중 일반적으로 자주 틀리는 발음이 포함되는지 여부를 식별할 수 있다. 상기 인식된 문자 중 자주 틀리는 발음이 존재하는 경우, 편차 모델(143)은 자주 틀리는 발음을 정확한 발음의 음소로 제시 및 변환할 수 있다.
- [0064] 교정 확인부(145)는 교정되는 발음에 대해 교정 전후 결과에 대한 단어 단위 또는 음소 단위 기반의 시각적인 결과를 제공할 수 있다.
- [0065] 음성 합성부(140-5)는 교정된 발음 결과 및 추출된 발화자의 목소리 특징을 음성 합성 모델을 이용하여 합성할 수 있다. 즉, 음성 합성부(140-5)는 발음 교정부(140-4)를 통해 교정된 음소 단위의 발음 및 화자 특징 추출부를 통해 출력된 사용자의 목소리 특징을 음성 합성 모델에 입력하여 사용자의 교정된 음성을 합성할 수 있다.
- [0067] 도 3은 본 개시의 일 실시예에 따른, 영어 원어민 수준의 발음 교정이 가능한 음성 변환 동작을 수행하는 방법

을 설명하기 위한 순서도이다.

- [0068] 장치는 사용자 음성을 마이크 등을 통해 입력받을 수 있다(S310).
- [0069] 장치는 사용자의 음성으로부터 사용자의 음성과 관련된 특징 데이터를 획득할 수 있다(S320).
- [0070] 일 예로, 장치는 사용자의 음성에 대한 멜-스펙트로그램(mel-spectrogram)을 CNN(convolution neural network) 모델 또는 어텐션 알고리즘 기반의 모델 중의 적어도 하나에 입력하여 사용자의 음성과 관련된 특징 데이터(예로, 사용자의 음성 스타일을 나타내는 특징 데이터 등)를 획득할 수 있다.
- [0071] 장치는 사용자의 음성에 대응되는 음소 단위의 문자를 획득할 수 있다(S330). 즉, 장치는 단어 등의 단위가 아닌 음소 단위로 사용자의 음성에 대응되는 문자를 획득할 수 있다.
- [0072] 장치는 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교를 통해 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 식별 및 교정할 수 있다(S340).
- [0073] 구체적으로, 장치는 음소 단위의 문자 및 발음 교정의 목표가 되는 목표 문자 간의 비교함으로써 교정이 필요한 제1 적어도 하나의 음소를 검출할 수 있다. 그리고, 장치는 제1 적어도 하나의 음소를 목표 문자를 구성하는 음소 형태로 치환/변환할 수 있다.
- [0074] 추가적으로, 장치는 음소 단위의 문자를 편차 모델에 입력하여 음소 단위의 문자 중 교정이 필요한 제2 적어도 하나의 음소를 식별 및 교정할 수 있다.
- [0075] 여기서, 상술한 바와 같이, 제1 적어도 하나의 음소는 상술한 바와 같이 사용자 음성에 대응되는 음소 단위의 문자와 목표 문자를 서로 비교함에 따라 검출된 교정이 필요한 음소를 의미한다. 제2 적어도 하나의 음소는 상기 음소 단위의 문자 중 편차 모델을 통해 교정이 필요한 것으로 검출된 음소를 의미한다. 여기서, 제1 적어도 하나의 음소 및 제2 적어도 하나의 음소 각각에서 중복되는 음소가 존재할 수도 있으나 이에 제한되는 것은 아니다.
- [0076] 장치는 목표 문자를 이용하여 사용자의 음성에 대응되는 음소 단위의 문자 중 교정이 필요한 제1 적어도 하나의 음소를 검출하고, 편차 모델을 통해 해당 음소 단위의 문자 중 교정이 필요한 제2 적어도 하나의 음소를 순차적으로 검출할 수 있다.
- [0077] 이에 따라, 장치는 목표 문자 및 사용자의 음성에 대응되는 음소 단위의 문자를 비교하는 절차를 통해 검출하지 못한 교정이 필요한 음소를 편차 모델을 통해 검출할 수도 있다. 즉, 상술된 장치의 동작을 통해 교정이 필요한 음소의 검출 확률을 높일 수 있다. 상술한 바와 같이, 편차 모델은 복수의 음소 유형에 대한 정답 발음 및 복수의 음소 유형에 대해 복수의 사용자의 실제 발음에 기초하여, 음소 단위의 문자 중 복수의 음소 유형에 대응되는 제2 적어도 하나의 음소를 식별하도록 학습될 수 있다.
- [0078] 즉, 장치는 편차 모델을 통해 음소 단위의 문자 중 일반적으로 많이 틀리는 것으로 예상되는 음소 유형이 존재하는지 식별할 수 있다. 음소 단위의 문자 중 일반적으로 많이 틀리는 것으로 예상되는 음소 유형에 대응되는 제2 적어도 하나의 음소가 식별된 경우, 장치는 제2 적어도 하나의 음소를 정답 발음에 대응되도록 교정할 수 있다.
- [0079] 장치는 교정 단계에 의해 교정된 음성 및 사용자의 음성과 관련된 특징 데이터에 기초하여, 최종 사용자 음성을 합성할 수 있다(S350).
- [0080] 구체적으로, 장치는 사용자의 음성과 관련된 특징 데이터, 교정된 제1 적어도 하나의 음소의 발음 및 제2 적어도 하나의 음소의 발음을 음성 합성 모델에 입력하여 최종 사용자 음성을 합성 및 획득할 수 있다.
- [0081] 장치는 최종 사용자 음성을 출력할 수 있다(S360). 일 예로, 장치는 최종 사용자 음성을 음성 형태로 출력할 수 있다.
- [0082] 이 때, 장치는 음소 단위의 문자로부터 교정된 제1 적어도 하나의 음소의 발음 및 제2 적어도 하나의 음소의 발음을 나타내는 정보를 시각적으로 표시할 수 있다.
- [0084] 이상, 본 개시에서 설명되는 실시예와 첨부된 도면은 본 개시에 포함되는 기술적 사상의 일부를 예시적으로 설명하는 것에 불과하다. 따라서, 본 명세서에 개시된 실시예들은 본 개시의 기술적 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이므로, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것이 아님은

자명하다.

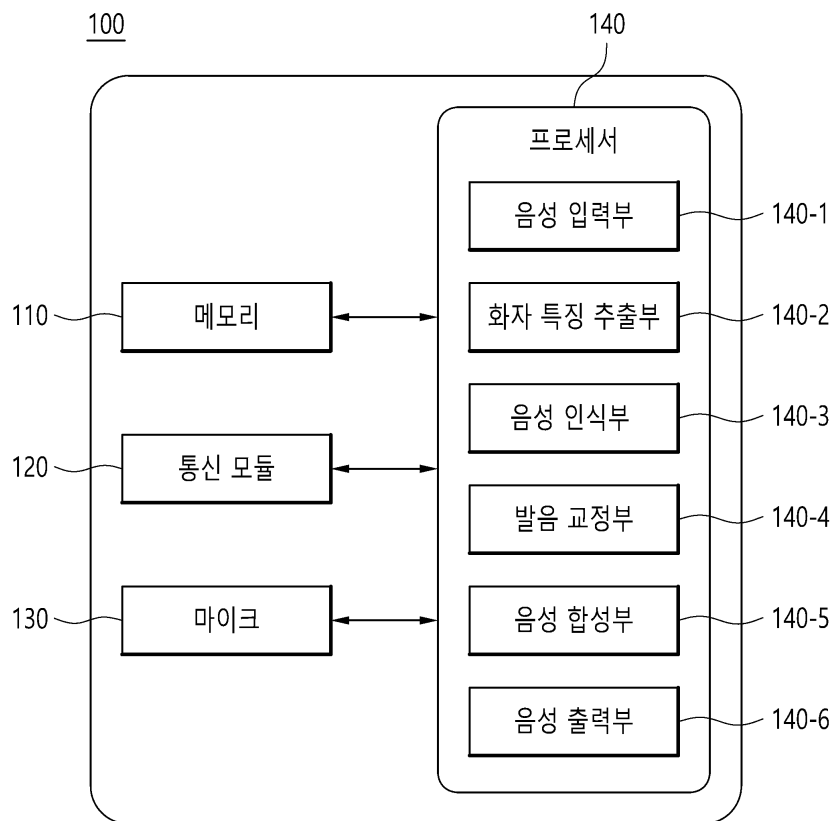
[0085] 본 개시의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시 예는 모두 본 개시의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

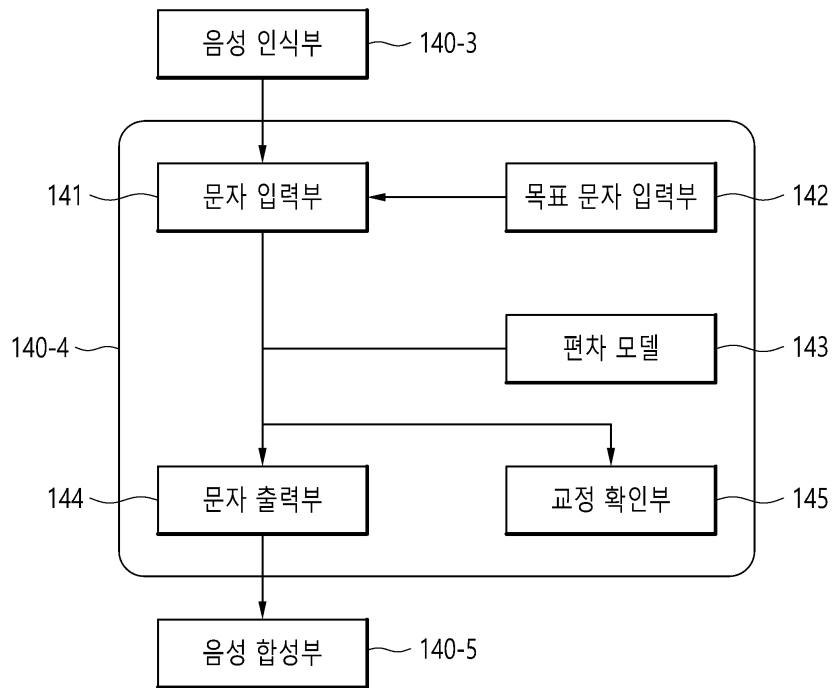
[0086] 100 : 장치
110 : 메모리
120 : 통신 모듈
140 : 프로세서
140-1 : 사전 학습부
140-2 : 미세 조정 학습부
140-3 : 추론부

도면

도면1



도면2



도면3

