(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0111884
(43) 공개일자 2023년07월26일

(51) 국제특허분류(Int. Cl.)
G10L 21/003 (2013.01) *G06N 3/08* (2023.01)
G10L 13/033 (2013.01) *G10L 13/08* (2006.01)
G10L 15/16 (2006.01) *G10L 19/26* (2013.01)
G10L 25/51 (2013.01)
(52) CPC특허분류
G10L 21/003 (2013.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2022-0007882
(22) 출원일자 2022년01월19일
심사청구일자 2022년01월19일

(71) 출원인
한림대학교 산학협력단
강원도 춘천시 한림대학길 1, 한림대학교(옥천동)
(72) 발명자
이은주
강원도 춘천시 국사봉길 7, 104동 1403호 (퇴계동)
장민욱
경기도 화성시 동탄반석로 231, 152동 2002호(석우동, 동탄예당마을 롯데캐슬)
(뒷면에 계속)
(74) 대리인
특허법인본

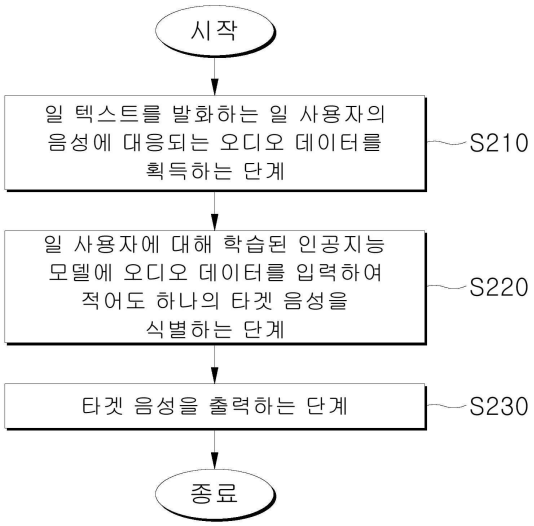
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **딥러닝 기반 구음 장애 음성 개선 변환 장치, 시스템의 제어 방법, 및 컴퓨터 프로그램**

(57) 요 약

본 발명은, 타겟 음성을 생성하기 위한 인공지능 모델이 저장된, 메모리, 및 메모리와 연결된 프로세서를 포함하는 전자 장치를 제공하고, 이때, 프로세서는, 일 텍스트를 발화하는 사용자의 제1 음성의 오디오 데이터와 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 바탕으로 인공지능 모델을 훈련시키고, 사용자의 제2 음성이 입력되면, 입력된 제2 음성의 오디오 데이터를 인공지능 모델에 입력하여, 인공지능 모델의 출력을 기반으로 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G10L 13/033 (2013.01)
G10L 13/08 (2013.01)
G10L 15/16 (2013.01)
G10L 19/26 (2013.01)
G10L 25/51 (2013.01)
(72) 발명자
김동욱
경기도 화성시 동탄공원로 21-12, 205동 2102호(능동, 포스코 더샵 아파트)
이동진
서울특별시 강남구 선릉로 120, 7동 301호(대치동, 개포우성아파트)
이재희
경기도 용인시 수지구 동천로153번길 6, 1412동 603호(동천동, 한빛마을 래미안 이스트팰리스 4단지)

명 세 서

청구범위

청구항 1

전자 장치에 있어서,

타겟 음성을 생성하기 위한 인공지능 모델이 저장된, 메모리; 및

상기 메모리와 연결된 프로세서;를 포함하고,

상기 프로세서는,

일 텍스트를 발화하는 사용자의 제1 음성의 오디오 데이터와 상기 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 바탕으로 상기 인공지능 모델을 훈련시키고,

상기 사용자의 제2 음성이 입력되면, 상기 입력된 제2 음성의 오디오 데이터를 상기 인공지능 모델에 입력하여, 상기 인공지능 모델의 출력을 기반으로 상기 사용자의 제2 음성이 변환된 타겟 음성을 획득하는, 전자 장치.

청구항 2

제1 항에 있어서,

상기 인공지능 모델은,

상기 제1 음성의 오디오 데이터로부터 추출된, 복수의 제1 특징 벡터와, 상기 제2 음성의 오디오 데이터로부터 추출된 복수의 제2 특징 벡터를 비교하여, 상기 제2 음성의 발화 대상인 텍스트를 식별하고,

상기 식별된 텍스트에 매칭되도록, 상기 제2 음성이 변환된 타겟 음성을 획득하는, 전자 장치.

청구항 3

제1 항에 있어서,

상기 프로세서는,

상기 제1 음성에 대한 오디오 데이터의 주파수 대역을 사용자 표준 데이터로 획득하고,

상기 사용자 표준 데이터의 주파수 대역을 기반으로, 상기 사용자의 제2 음성이 변환된 타겟 음성을 획득하는, 전자 장치.

청구항 4

제1 항에 있어서,

상기 프로세서는,

상기 제2 음성의 발화 대상인 텍스트에 대한 제2 음성의 발음 정확도를 판단하고,

상기 발음 정확도가 임계치 이상인 경우, 상기 제2 음성을 타겟 음성으로 출력하고,

상기 발음 정확도가 임계치 미만인 경우, 상기 인공지능 모델의 출력을 기반으로 상기 사용자의 제2 음성이 변환된 타겟 음성을 획득하는, 전자 장치.

청구항 5

제4 항에 있어서,

상기 프로세서는,

상기 제2 음성의 발화 대상인 텍스트를 구성하는 제1 서브 텍스트 및 제2 서브 텍스트를 획득하고,

상기 제2 음성의 오디오 데이터 중 상기 제1 서브 텍스트 및 상기 제2 서브 텍스트 각각이 매칭된 오디오 구간

을 식별하고,

상기 제1 서브 텍스트에 대한 제1 오디오 구간의 발음 정확도 및 상기 제2 서브 텍스트에 대한 제2 오디오 구간의 발음 정확도를 식별하고,

상기 제1 서브 텍스트에 대한 발음 정확도가 임계치 미만이고, 상기 제2 서브 텍스트에 대한 발음 정확도가 임계치 이상이면, 상기 제1 서브 텍스트가 입력된 상기 인공지능 모델의 출력을 기반으로 상기 제1 서브 텍스트에 대한 타겟 음성을 획득하고,

상기 제1 서브 텍스트에 대한 타겟 음성 및 상기 제2 오디오 구간을 포함하는 오디오 데이터를 바탕으로, 제2 음성이 변환된 음성을 출력하는, 전자 장치.

청구항 6

제4 항에 있어서,

상기 프로세서는,

복수의 텍스트 각각에 대한 상기 사용자의 음성을 획득하여, 음성 인식을 실시함으로써 상기 복수의 텍스트 각각에 대한 발음 정확도를 매칭하여 저장하고,

상기 복수의 텍스트 중, 기 매칭된 발음 정확도가 임계치 미만인 제1 텍스트를 발화하는, 상기 사용자의 새로 입력된 음성의 발음 정확도가 임계치 이상인 경우, 상기 제1 텍스트를 발화하는 사용자의 음성인, 제1 테스트 음성을 일정 횟수만큼 획득하고,

상기 일정 횟수만큼 획득된 제1 테스트 음성의 발음 정확도에 대한 평균이 임계치 이상이면, 상기 제1 테스트 음성의 발음 정확도에 대한 평균을 바탕으로 상기 제1 텍스트에 매칭된 발음 정확도를 업데이트하고,

상기 복수의 텍스트 중, 기 매칭된 발음 정확도가 임계치 이상인 제2 텍스트를 발화하는, 상기 사용자의 새로 입력된 음성의 발음 정확도가 임계치 미만인 경우, 상기 제2 텍스트를 발화하는 사용자의 음성인, 제2 테스트 음성을 일정 횟수만큼 획득하고,

상기 일정 횟수만큼 획득된 테스트 음성의 발음 정확도에 대한 평균이 임계치 미만이면, 상기 제2 테스트 음성의 발음 정확도에 대한 평균을 바탕으로 상기 제2 텍스트에 매칭된 발음 정확도를 업데이트하는, 전자 장치.

청구항 7

제1 항에 있어서,

상기 프로세서는,

오디오 데이터를 노이즈 필터링(noise filtering)하고,

상기 필터링된 오디오 데이터를 복수의 구간으로 구간화하되, 인접한 구간 간에 일정 시간만큼 중첩하여 프레이밍(framing)하고,

상기 프레이밍된 오디오 데이터로부터 복수의 음향(aqoustic) 특징 벡터 및 복수의 MFCC(Mel Frequency Cepstral Coefficients) 특징 벡터를 추출하는 것을 특징으로 하는, 전자 장치.

청구항 8

제7 항에 있어서,

상기 프로세서는,

상기 프레이밍된 오디오 데이터에 MFCC 기법을 적용하여 적어도 하나의 제1 MFCC 특징 벡터를 획득하고,

상기 적어도 하나의 제1 MFCC 특징 벡터의 값을 미분하여 적어도 하나의 제2 MFCC 특징 벡터를 획득하고,

상기 적어도 하나의 제2 MFCC 특징 벡터의 값을 미분하여 적어도 하나의 제3 MFCC 특징 벡터를 획득하며,

상기 복수의 MFCC 특징 벡터는,

상기 적어도 하나의 제1 MFCC 특징 벡터, 상기 적어도 하나의 제2 MFCC 특징 벡터, 및 상기 적어도 하나의 제3

MFCC 특징 벡터를 포함하는 것을 특징으로 하는, 전자 장치.

청구항 9

서버에 의한 딥러닝 기반 구음 장애 음성 개선 변환 시스템의 제어 방법에 있어서,

일 사용자 단말로부터, 일 텍스트를 발화하는 사용자의 제1 음성의 오디오 데이터를 획득하는 단계;

상기 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 획득하는 단계;

상기 제1 음성의 오디오 데이터 및 상기 정상적인 음성의 오디오 데이터를 바탕으로, 타겟 음성을 출력하도록 인공지능 모델을 훈련시키는 단계;

상기 사용자 단말로부터 상기 사용자의 제2 음성을 수신하면, 상기 수신된 제2 음성의 오디오 데이터를 상기 인공지능 모델에 입력하는 단계;

상기 인공지능 모델로부터 상기 제2 음성이 변환된 타겟 음성을 획득하는 단계; 및

상기 사용자 단말을 통해, 상기 제2 음성이 변환된 타겟 음성을 출력하는 단계;를 포함하는, 시스템의 제어 방법.

청구항 10

제9 항에 있어서,

상기 제2 음성이 변환된 타겟 음성을 획득하는 단계는,

상기 제1 음성의 오디오 데이터로부터 추출된, 복수의 제1 특징 벡터와, 상기 제2 음성의 오디오 데이터로부터 추출된 복수의 제2 특징 벡터를 비교하여, 상기 제2 음성의 발화 대상인 텍스트를 식별하는 단계; 및

상기 식별된 텍스트에 매칭되도록, 상기 제2 음성이 변환된 타겟 음성을 획득하는 단계;를 포함하고,

복수의 특징 벡터는,

일 오디오 데이터로부터 추출된 복수의 음향(aqoustic) 특징 벡터 및 복수의 MFCC(Mel Frequency Cepstral Coefficients) 특징 벡터를 포함하는, 시스템의 제어 방법.

청구항 11

제9 항에 있어서,

상기 타겟 음성을 출력하도록 인공지능 모델을 훈련시키는 단계는,

상기 제1 음성에 대한 오디오 데이터의 주파수 대역을 사용자 표준 데이터로 획득하는 단계; 및

상기 사용자 표준 데이터의 주파수 대역을 기반으로, 상기 사용자의 제2 음성이 변환된 타겟 음성을 획득하는 단계;를 포함하는, 시스템의 제어 방법.

청구항 12

하드웨어인 컴퓨터와 결합되어, 제9 항의 방법을 수행할 수 있도록 컴퓨터에서 독출가능한 기록매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 일 텍스트를 발화하는 환자의 음성이 입력되는 즉시, 환자의 음성이 변환된, 정상적인 음성인 타겟 음성을 출력하기 위한 딥러닝 기반 구음 장애 음성 개선 변환 기술에 관한 것으로, 타겟 음성은 환자의 선택에 따라, 본인의 음성, 남성의 음성, 여성의 음성, 중년인의 음성, 특정인의 음성 등 다양한 음성으로 변환되어 출력될 수 있다.

배 경 기 술

[0002] 구음 장애를 갖는 환자의 경우, 일상생활에서의 의사소통이 어려우며, 이 경우, 평가를 통해 재활을 실시할 수 있으나, 시간적 제약으로 인해, 당장 의사소통이 필요한 상황에 도움이 되지는 못한다는 문제점이 존재한다.

[0003] 이와 관련하여, 특허문헌 1과 같은 구음 장애에 대한 평가를 실시하는 기술은, 평가를 실시하는데 그칠뿐이며, 복잡한 절차에 따른 언어 능력 테스트 수행 및 설문지의 작성은 시간적 제약이 존재하여 비효율적으로 이루어졌다.

[0004] 한편, 상기의 배경기술로서 설명된 사항들은 본 발명의 배경에 대한 이해 증진을 위한 것일 뿐, 이 기술분야에서 통상의 지식을 가진 자에게 이미 알려진 종래기술에 해당함을 인정하는 것으로 받아들여져서는 안 될 것이다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 등록특허공보 제10-1921890호, 2018.11.20.

발명의 내용

해결하려는 과제

[0006] 본 발명이 해결하고자 하는 과제는 실시간으로 의사 소통을 수행하는데 있어서, 구음 장애를 갖는 환자의 음성을 정상적인 음성으로 출력하는 딥러닝 기반 구음 장애 음성 개선 변환 기술을 제공하는 것이다.

[0007] 본 발명의 또 다른 목적으로는, 환자가 입력한 음성을 정상적인 음성인 타겟 음성으로 출력하여 제공함으로써 즉석에서 구음 장애 개선을 위한 재활 훈련을 실시할 수 있도록 하는 딥러닝 기반 구음 장애 음성 개선 변환 기술을 제공하는 것이다.

[0008] 또한 본 발명은, 한국지능정보사회진흥원, 인공지능 학습용 데이터 구축사업의 구음장애 음성 데이터(과제고유번호 2100213130508821210010 1BS)로 2021.06.01부터 2021.12.31의 기간에 연구된 기술이다.

[0009] 본 발명이 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상술한 과제를 해결하기 위한 본 발명의 일 면에 따른 전자 장치에 있어서, 타겟 음성을 생성하기 위한 인공지능 모델이 저장된, 메모리, 및 메모리와 연결된 프로세서를 포함한다.

[0011] 여기서, 프로세서는, 일 텍스트를 발화하는 사용자의 제1 음성의 오디오 데이터와 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 바탕으로 인공지능 모델을 훈련시키고, 사용자의 제2 음성이 입력되면, 입력된 제2 음성의 오디오 데이터를 인공지능 모델에 입력하여, 인공지능 모델의 출력을 기반으로 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0012] 또한, 인공지능 모델은, 제1 음성의 오디오 데이터로부터 추출된, 복수의 제1 특징 벡터와, 제2 음성의 오디오 데이터로부터 추출된 복수의 제2 특징 벡터를 비교하여, 제2 음성의 발화 대상인 텍스트를 식별하고, 식별된 텍스트에 매칭되도록, 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0013] 이때, 프로세서는, 제1 음성에 대한 오디오 데이터의 주파수 대역을 사용자 표준 데이터로 획득하고, 사용자 표준 데이터의 주파수 대역을 기반으로, 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0014] 추가로, 프로세서는, 제2 음성의 발화 대상인 텍스트에 대한 제2 음성의 발음 정확도를 판단하고, 발음 정확도가 임계치 이상인 경우, 제2 음성을 타겟 음성으로 출력하고, 발음 정확도가 임계치 미만인 경우, 인공지능 모델의 출력을 기반으로 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0015] 바람직하게는, 프로세서는, 제2 음성의 발화 대상인 텍스트를 구성하는 제1 서브 텍스트 및 제2 서브 텍스트를 획득하고, 제2 음성의 오디오 데이터 중 제1 서브 텍스트 및 제2 서브 텍스트 각각이 매칭된 오디오 구간을 식별한다.

[0016] 이후, 프로세서는, 제1 서브 텍스트에 대한 제1 오디오 구간의 발음 정확도 및 제2 서브 텍스트에 대한 제2 오디오 구간의 발음 정확도를 식별하고, 제1 서브 텍스트에 대한 발음 정확도가 임계치 미만이고, 제2 서브 텍스트에 대한 발음 정확도가 임계치 이상이면, 제1 서브 텍스트가 입력된 인공지능 모델의 출력을 기반으로 제1 서브 텍스트에 대한 타겟 음성을 획득한다.

[0017] 이에 따라, 프로세서는, 제1 서브 텍스트에 대한 타겟 음성 및 제2 오디오 구간을 포함하는 오디오 데이터를 바탕으로, 제2 음성이 변환된 음성을 출력할 수 있다.

[0018] 이를 위해, 프로세서는, 복수의 텍스트 각각에 대한 사용자의 음성을 획득하여, 음성 인식을 실시함으로써 복수의 텍스트 각각에 대한 발음 정확도를 매칭하여 저장한다.

[0019] 한편, 프로세서는, 복수의 텍스트 중, 기 매칭된 발음 정확도가 임계치 미만인 제1 텍스트를 발화하는, 사용자의 새로 입력된 음성의 발음 정확도가 임계치 이상인 경우, 제1 텍스트를 발화하는 사용자의 음성인, 제1 테스트 음성을 일정 횟수만큼 획득하고, 일정 횟수만큼 획득된 제1 테스트 음성의 발음 정확도에 대한 평균이 임계치 이상이면, 제1 테스트 음성의 발음 정확도에 대한 평균을 바탕으로 제1 텍스트에 매칭된 발음 정확도를 업데이트할 수 있다.

[0020] 반면, 프로세서는, 복수의 텍스트 중, 기 매칭된 발음 정확도가 임계치 이상인 제2 텍스트를 발화하는, 사용자의 새로 입력된 음성의 발음 정확도가 임계치 미만인 경우, 제2 텍스트를 발화하는 사용자의 음성인, 제2 테스트 음성을 일정 횟수만큼 획득하고, 일정 횟수만큼 획득된 테스트 음성의 발음 정확도에 대한 평균이 임계치 미만이면, 제2 테스트 음성의 발음 정확도에 대한 평균을 바탕으로 제2 텍스트에 매칭된 발음 정확도를 업데이트할 수 있다.

[0021] 추가로, 프로세서는, 오디오 데이터를 노이즈 필터링(noise filtering)하고, 필터링된 오디오 데이터를 복수의 구간으로 구간화하되, 인접한 구간 간에 일정 시간만큼 중첩하여 프레이밍(framing)하고, 프레이밍된 오디오 데이터로부터 복수의 음향(aqoustic) 특징 벡터 및 복수의 MFCC(Mel Frequency Cepstral Coefficients) 특징 벡터를 추출하는 것을 특징으로 할 수 있다.

[0022] 이때, 프로세서는, 프레이밍된 오디오 데이터에 MFCC 기법을 적용하여 적어도 하나의 제1 MFCC 특징 벡터를 획득하고, 적어도 하나의 제1 MFCC 특징 벡터의 값을 미분하여 적어도 하나의 제2 MFCC 특징 벡터를 획득하고, 적어도 하나의 제2 MFCC 특징 벡터의 값을 미분하여 적어도 하나의 제3 MFCC 특징 벡터를 획득하며, 복수의 MFCC 특징 벡터는, 적어도 하나의 제1 MFCC 특징 벡터, 적어도 하나의 제2 MFCC 특징 벡터, 및 적어도 하나의 제3 MFCC 특징 벡터를 포함하는 것을 특징으로 할 수 있다.

[0023] 한편, 본 발명의 이 면에 따른 서버에 의한 딥러닝 기반 구음 장애 음성 개선 변환 시스템의 제어 방법에 있어서, 일 사용자 단말로부터, 일 텍스트를 발화하는 사용자의 제1 음성의 오디오 데이터를 획득하는 단계, 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 획득하는 단계, 제1 음성의 오디오 데이터 및 정상적인 음성의 오디오 데이터를 바탕으로, 타겟 음성을 출력하도록 인공지능 모델을 훈련시키는 단계, 사용자 단말로부터 사용자의 제2 음성을 수신하면, 수신된 제2 음성의 오디오 데이터를 인공지능 모델에 입력하는 단계, 인공지능 모델로부터 제2 음성이 변환된 타겟 음성을 획득하는 단계, 및 사용자 단말을 통해, 제2 음성이 변환된 타겟 음성을 출력하는 단계를 포함할 수 있다.

[0024] 이때, 제2 음성이 변환된 타겟 음성을 획득하는 단계는, 제1 음성의 오디오 데이터로부터 추출된, 복수의 제1 특징 벡터와, 제2 음성의 오디오 데이터로부터 추출된 복수의 제2 특징 벡터를 비교하여, 제2 음성의 발화 대상인 텍스트를 식별하는 단계, 및 식별된 텍스트에 매칭되도록, 제2 음성이 변환된 타겟 음성을 획득하는 단계를 포함하고, 복수의 특징 벡터는, 일 오디오 데이터로부터 추출된 복수의 음향(aqoustic) 특징 벡터 및 복수의 MFCC(Mel Frequency Cepstral Coefficients) 특징 벡터를 포함할 수 있다.

[0025] 바람직하게, 타겟 음성을 출력하도록 인공지능 모델을 훈련시키는 단계는, 제1 음성에 대한 오디오 데이터의 주파수 대역을 사용자 표준 데이터로 획득하는 단계, 및 사용자 표준 데이터의 주파수 대역을 기반으로, 사용자의 제2 음성이 변환된 타겟 음성을 획득하는 단계를 포함할 수 있다.

[0026] 한편, 본 발명은 하드웨어인 컴퓨터와 결합되어, 본 발명의 이 면에 따른 방법을 수행할 수 있도록 컴퓨터에서 독출가능한 기록매체에 저장된 컴퓨터 프로그램을 포함한다.

[0027] 본 발명의 기타 구체적인 사항들은 상세한 설명 및 도면들에 포함되어 있다.

발명의 효과

[0028] 본 발명의 딥러닝 기반 구음 장애 음성 개선 변환 기술에 의하면, 사용자의 음성 입력만으로도 사용자가 발화하는 텍스트를 식별하고, 텍스트에 대한 정상적인 음성을 타겟 음성으로 출력하여 제공함으로써, 구음 장애를 갖는 환자가 실시간 의사소통을 정확하고 빠르게 진행할 수 있도록 한다.

[0029] 본 발명의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0030] 도 1은 본 발명의 일 실시예에 따른 전자 장치의 구성도이다.

도 2는 본 발명의 일 실시예에 따른 딥러닝 기반 구음 장애 음성 개선 변환 기본 흐름도이다.

도 3 및 도 4는 본 발명의 일 실시예에 따른 다양한 시스템 구성도이다.

도 5는 본 발명의 일 실시예에 따른 시스템의 동작 흐름도이다.

도 6은 본 발명의 일 실시예에 따른 음성 개선 변환 실험 결과도이다.

도 7은 본 발명의 이 실시예에 따른 사용자 단말을 통한 실생활 사용 예시도이다.

도 8은 본 발명의 일 실시예에 다른 서버 구성도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야의 통상의 기술자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0032] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 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. 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 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0033] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 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.

[0034] 명세서에서 사용되는 "부" 또는 “모듈”이라는 용어는 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소를 의미하며, "부" 또는 “모듈”은 어떤 역할들을 수행한다. 그렇지만 "부" 또는 “모듈”은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부" 또는 “모듈”은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 "부" 또는 “모듈”은 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 "부" 또는 “모듈”들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 "부" 또는 “모듈”들로 결합되거나 추가적인 구성요소들과 "부" 또는 “모듈”들로 더 분리될 수 있다.

[0035] 공간적으로 상대적인 용어인 "아래(below)", "아래(beneath)", "하부(lower)", "위(above)", "상부(upper)" 등은 도면에 도시되어 있는 바와 같이 하나의 구성요소와 다른 구성요소들과의 상관관계를 용이하게 기술하기 위

해 사용될 수 있다. 공간적으로 상대적인 용어는 도면에 도시되어 있는 방향에 더하여 사용시 또는 동작시 구성요소들의 서로 다른 방향을 포함하는 용어로 이해되어야 한다. 예를 들어, 도면에 도시되어 있는 구성요소를 뒤집을 경우, 다른 구성요소의 "아래(below)"또는 "아래(beneath)"로 기술된 구성요소는 다른 구성요소의 "위(above)"에 놓일 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.

[0036] 본 명세서에서, 컴퓨터는 적어도 하나의 프로세서를 포함하는 모든 종류의 하드웨어 장치를 의미하는 것이고, 실시 예에 따라 해당 하드웨어 장치에서 동작하는 소프트웨어적 구성도 포괄하는 의미로서 이해될 수 있다. 예를 들어, 컴퓨터는 스마트폰, 태블릿 PC, 데스크톱, 노트북 및 각 장치에서 구동되는 사용자 클라이언트 및 애플리케이션을 모두 포함하는 의미로서 이해될 수 있으며, 또한 이에 제한되는 것은 아니다.

[0037] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0038] 도 1은 본 발명의 일 실시예에 따른 전자 장치의 구성도이다.

[0039] 도시된 바와 같이, 본 발명의 일 면에 따른 전자 장치(100)는, 타겟 음성을 생성하기 위한 인공지능 모델(111)이 저장된, 메모리(110), 및 메모리(110)와 연결된 프로세서(120)를 포함한다.

[0040] 일 실시예로, 전자 장치(100)는 스마트폰(smartphone), 태블릿 PC(tablet personal computer), 이동 전화기(mobile phone), 영상 전화기, 전자책 리더기(e-book reader), 데스크탑 PC (desktop PC), 랩탑 PC(laptop PC), 넷북 컴퓨터(netbook computer), 워크스테이션(workstation), 서버, PDA(personal digital assistant), PMP(portable multimedia player), MP3 플레이어, 모바일 의료기기, 카메라, 또는 웨어러블 장치(wearable device), 인공지능 스피커(AI speaker) 중 적어도 하나를 포함할 수 있다.

[0041] 도 2는 본 발명의 일 실시예에 따른 딥러닝 기반 구음 장애 음성 개선 변환 기본 흐름도이다.

[0042] 프로세서(120)는, 일 텍스트를 발화하는 일 사용자의 제1 음성의 오디오 데이터와 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 바탕으로 일 사용자에 대해 학습된 인공지능 모델(111)을 통해, 딥러닝 기반 구음 장애 음성 개선 변환을 실행할 수 있다.

[0043] 구체적으로, 도 2에 도시된 바와 같이, 일 사용자의 제2 음성이 입력되면, 입력된 제2 음성의 오디오 데이터(음성으로부터 추출되는 디지털화된 음원)를 획득(S210)하고, 이를 일 사용자의 음성에 대해 학습된 인공지능 모델(111)에 입력하여, 인공지능 모델(111)의 출력을 기반으로 사용자의 제2 음성이 변환된 타겟 음성을 획득(S220)하여 사용자에게 제공(S230)할 수 있다.

[0044] 이를 위해, 인공지능 모델(111)은, 제1 음성의 오디오 데이터로부터 추출된, 복수의 제1 특징 벡터와, 제2 음성의 오디오 데이터로부터 추출된 복수의 제2 특징 벡터를 비교하여, 제2 음성의 발화 대상인 텍스트를 식별하고, 식별된 텍스트에 매칭되도록, 제2 음성이 변환된 타겟 음성을 획득한다.

[0045] 이 경우, 각 텍스트 별로, 사용자의 음성(: 제1 음성)의 복수의 특징 벡터의 값이 저장될 수 있다. 이때, 각 텍스트 별로 저장된 복수의 특징 벡터를 활용하여, 적어도 하나의 제1 음성 인식 모델이 훈련될 수도 있다. 그 결과, 제1 음성 인식 모델은, 사용자(ex. 구음 장애 환자)에게 맞춤형으로 훈련되어, 사용자가 발화한 텍스트를 인식할 수 있다.

[0046] 한편, 프로세서(120)는 구음 장애를 갖는 사용자의 음성이 입력되면, 음성에 따른 텍스트를 식별하고, 식별된 텍스트에 대해 기 저장된 성우 음성, 또는 TTS(Text-To-Speech) 기술에 의해 생성된 음성 등의 정상적인 음성을 타겟 음성으로 출력할 수 있다.

[0047] 구체적으로, 타겟 음성을 획득하는데 있어서, 프로세서(120)는 제1 음성에 대한 오디오 데이터의 주파수 대역을 사용자 표준 데이터로 획득하고, 사용자 표준 데이터의 주파수 대역을 기반으로, 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0048] 이때, 프로세서(120)는, 일 텍스트에 대한 복수의 타겟 음성을 획득할 수 있으며, 복수의 타겟 음성은 각각 주파수 대역이 다르게 설정되어, 다른 음색으로 출력되도록 한다.

[0049] 이에 따라, 복수의 타겟 음성은, 남성 음색, 여성 음색, 노인 음색, 어린이 음색, 특정 인물 음색, 및 사용자 표준 데이터에 기반한 사용자 음색 등에 따라 구성될 수 있다.

[0050] 한편, 구음 장애는, 후천적 또는 선천적인 요인에 의한 장애 및/또는 질병에 의해 발성 기관에 생긴 기능 이상으로 말하기 어려운 상태가 된 것을 뜻하며, 구음 장애 유형으로는, 구마비, 가성구마비, 청각 장애, 후두 장애, 언어 장애, 파킨슨병 등이 포함된다.

[0051] 도 3 및 도 4는 본 발명의 일 실시예에 따른 다양한 시스템 구성도이고, 도 5는 본 발명의 일 실시예에 따른 시스템의 동작 흐름도이다.

[0052] 도 3에 도시된 바와 같이, 본 발명의 전자 장치(100)는, 딥러닝 기반 구음 장애 시각화 및 재활을 수행하는데 있어서, 음성 입출력부(130)를 포함하여, 음성 입출력부(130)의 마이크(130a)로부터 사용자의 오디오 데이터를 획득하고, 음성 입출력부(130)의 스피커(130b)를 통해 제2 음성이 변환된 타겟 음성을 출력할 수 있다.

[0053] 이때, 전자 장치(100)는 디스플레이(140)를 포함하여, 사용자가 원하는 타겟 음성의 음색에 대한 UI의 선택에 대한 사용자 명령을 획득할 수 있다.

[0054] 또한, 도 4에 도시된 바와 같이, 본 발명의 전자 장치(100)는, 통신부(150)를 포함하여 사용자 단말(200)과 통신함으로써, 사용자 단말(200)을 통해 오디오 데이터를 획득하여, 제2 음성이 변환된 타겟 음성을 출력한다.

[0055] 이때, 도 5에 도시된 바와 같이, 사용자 단말(200)을 통해 사용자가 원하는 타겟 음성의 음색에 대한 UI의 선택에 대한 사용자 명령을 획득할 수 있다.

[0056] 이때, 사용자 단말은 스마트폰(smartphone), 태블릿 PC(tablet personal computer), 이동 전화기(mobile phone), 영상 전화기, 전자책 리더기(e-book reader), 데스크탑 PC (desktop PC), 랩탑 PC(laptop PC), 넷북 컴퓨터(netbook computer), 워크스테이션(workstation), 서버, PDA(personal digital assistant), PMP(portable multimedia player), MP3 플레이어, 모바일 의료기기, 카메라, 또는 웨어러블 장치(wearable device), 인공지능 스피커(AI speaker) 중 적어도 하나를 포함할 수 있다.

[0057] 도 5에 도시된 바와 같이, 복수의 특징 벡터를 획득하는데 있어서, 획득된 오디오 데이터는, 프로세서(120)에 의해, 노이즈 필터링(noise filtering) 작업을 거치며, 프로세서(120)는 필터링된 오디오 데이터를 복수의 구간으로 구간화하되, 인접한 구간 간에 일정 시간만큼 중첩하여 프레이밍(framing)하고, 프레이밍된 오디오 데이터로부터 복수의 음향(aqoustic) 특징 벡터 및 복수의 MFCC(Mel Frequency Cepstral Coefficients) 특징 벡터를 추출하며, 복수의 특징 벡터는, 복수의 음향 특징 벡터 및 복수의 MFCC 특징 벡터를 포함한다.

[0058] 이때, 복수의 음향 특징 벡터는, meanF0Hz, stdevF0Hz, meanF1Hz, stdevF1Hz, meanF2Hz, stdevF2Hz, HNR, localjitter, localabsolutejitter, rapjitter, ppq5jitter, ddpjitter, localShimmer, localdbShimmer, apq3Shimmer, apq5Shimmer, apq11Shimmer, ddaShimmer, pitch_max 및 pich_min 중 적어도 하나를 포함하여, 프레이밍된 오디오 데이터로부터 프로세서(120)에 의해 추출될 수 있다.

[0059] 이때, 복수의 MFCC 특징 벡터는, 적어도 하나의 제1 MFCC 특징 벡터, 적어도 하나의 제2 MFCC 특징 벡터, 및 적어도 하나의 제3 MFCC 특징 벡터를 포함한다.

[0060] 제1 MFCC 특징 벡터는, 프로세서(120)가 프레이밍된 오디오 데이터에 MFCC 기법을 적용함으로써 최대 13개의 벡터로 추출된다.

[0061] 제2 MFCC 특징 벡터는, 프로세서(120)에 의해, 적어도 하나의 제1 MFCC 특징 벡터의 값을 미분하여 추출되며, 제3 MFCC 특징 벡터는, 프로세서(120)에 의해, 적어도 하나의 제2 MFCC 특징 벡터의 값을 미분하여 추출될 수 있다.

[0062] 이에 따라, 프로세서(120)는, 음향 특징 벡터 최대 20개, MFCC 특징 벡터 최대 78개(각 프레임 마다, 제1 MFCC 특징 벡터의 최대 개수인 13개, 1회 미분으로 인한 최대 개수인 13개, 3회 미분으로 인한 최대 개수인 13개, 총 13*3=39개를 구한 다음 이들 각각을 모든 프레임에 대하여 구한 평균 및 표준편차를 구하여 특징 벡터로 사용, 따라서 하나의 오디오 데이터에 대하여 최대 13*3*2=78개의 MFCC 특징 벡터를 추출)를 포함하여 최대 98개의 특징 벡터를 오디오 데이터로부터 추출할 수 있다.

[0063] 프로세서(120)는 수집된 사용자의 음성에 대한 오디오 데이터를 학습용 데이터와 예측용 데이터를 일정 비율(ex. 학습용 8: 예측용 2)로 분류하여 학습용 데이터를 통해 인공지능 모델(111)을 학습시키고, 예측용 데이터를 학습된 인공지능 모델(111)에 입력하여 출력된 분류 결과를 바탕으로 인공지능 모델(111)의 정확도를 판단할 수 있다.

[0064] 이에 따라, 구음 장애가 발생한 음성을 타겟 음성으로 변환하는 인공지능 모델(111)의 정확도는 99% 이상을 목

표로 하여, 학습이 수행될 수 있다.

[0065] 한편, 인공지능 모델(111)은 타겟 음성 생성 모델을 포함할 수 있다.

[0066] 타겟 음성 생성 모델은 사용자 음성 신호인 소스 음성 신호, 사용자 음서의 발화 대상인 텍스트, 및 사용자 음성으로 변환한 타겟 음성에 대한 타겟 음성 신호를 입력받는 인코더를 포함한다.

[0067] 또한, 타겟 음성 생성 모델은 attention 기반 스펙트로그램(spectrogram) 디코더 및 자동음성인식(ASR) 디코더(보조적으로 추가 사용하는 디코더, 출력 타겟 음성 신호의 텍스트를 예측하는 과정을 통해 타겟 음성 생성 모델의 학습 결과에 대한 정확도를 향상시킴, attention 기반 LSTM)를 포함하며, 이때, 인코더와 디코더 사이에 attention mechanism을 사용함으로써 모델이 텍스트와 스펙트로그램 간의 관계를 학습하도록 한다.

[0068] 이에 따른, 타겟 음성 생성 모델은 스펙트로그램 디코더로부터 출력된 스펙트로그램을 Griffin-Lim 알고리즘(GLA, 손실된 phase 정보를 유추하여 선형 스펙트로그램을 음성으로 복원시키는 기술) 기반의 음성 합성기를 통해 변환하여 타겟 음성 신호를 생성할 수 있다.

[0069] 한편, 모델링 알고리즘에 있어서, end-to-end sequence-to-sequence 모델 구조를 적용할 수 있으며, 이때 역시, 입력 데이터는 사용자의 음성 신호이고, 출력 데이터는 음성 신호가 변환된 타겟 음성 신호이다.

[0070] 학습 과정에 있어서, 사용자 음성 신호인 소스 음성 신호, 사용자 음서의 발화 대상인 텍스트, 및 사용자 음성으로 변환한 타겟 음성에 대한 타겟 음성 신호 각각에 대한 데이터가 요구된다.

[0071] end-to-end sequence-to-sequence 모델 구조의 적용 예시로써, 신경망을 구성하는 인코더, 디코더 및 보코더는 각각 하나씩 적용되며, 이때, 디코더는 attention 기반 디코더이다.

[0072] 이에 따르면, 인코더는 음성 신호를 프레임별 히든 특징벡터로 생성하고, 디코더는 인코더로부터 출력된 데이터를 바탕으로, 음성 신호의 스펙트로그램을 출력하고, 보코더는 디코더의 출력을 바탕으로, 시간 영역의 wave 신호를 출력할 수 있다.

[0073] 구체적으로, 인코더는, 16kHz 단위로 샘플링된 음성 신호를 프로세서(120)로부터 입력받아, 512-dimentional 인코더 represntation을 출력함에 있어, 다음과 같은 과정을 수행할 수 있다.

[0074] 16kHz 단위로 샘플링된 음성 신호는 80 dimensional log-mel spectrogram features over a range of 125-7600Hz, 2개의 convolutional layers with ReLU activations(다운샘플링 by 4), 주파수 영역에 대해서만 convolving하는 하나의 bidirectional convolutional LSTM layers, 및 3개의 bidirectional LSTM layers를 차례로 통과하여, 출력 데이터인 512-dimentional 인코더 represntation을 출력할 수 있다.

[0075] 이후, 디코더는, 인코더의 출력을 입력받아, 이를 스펙트로그램을 한번에 한 프레임씩 예측하는 하나의 autoregressive RNN을 통과시킴으로써 타겟 스펙트로그램, 즉, 타겟 음성 신호를 출력할 수 있다.도 6은 본 발명의 일 실시예에 따른 음성 개선 변환 실험 결과도이다.

[0076] 프로세서(120)로의 일 텍스트(ex. 전화, 동물원, 자동차, 땅콩, 엄마, 귀, 짹짹, 병원 등)를 발화하는 환자의 음성의 입력에 대해, google TTS를 통해 유사 발음 또는 동일한 발음의 타겟 음성을 획득할 수 있었으며, 프로세서(120)에 의한 인공지능 모델(111)의 학습 횟수에 따라, Train loss가 감소함이 증명되었다.

[0077] 도 7은 본 발명의 이 실시예에 따른 사용자 단말(200)을 통한 실생활 사용 예시도이다.

[0078] 실시예로써, 환자의 보호자가 AI 스피커를 통해 상황에 따른 문장 리스트를 요청하면, AI 스피커로부터 상황에 따른 문장 리스트가 생성한다.

[0079] 이때, AI 스피커는, 환자의 제1 음성에 대해 학습된 인공지능 모델(111)을 통해 환자의 음색을 갖는 타겟 음성을 문장 리스트에 적용하여 환자의 단말로 전송할 수 있다.

[0080] 환자가 상황에 따른 의사소통을 실시할 때, 환자의 단말은 환자의 제2 음성 및 문장 리스트를 바탕으로, 별도의 변환 작업 없이 빠른 속도로 타겟 음성을 출력할 수 있다.

[0081] 한편, 실시예로써, 프로세서(120)는, 제2 음성의 발화 대상인 텍스트에 대한 제2 음성의 발음 정확도를 판단하고, 발음 정확도가 임계치 이상인 경우, 제2 음성을 타겟 음성으로 출력하고, 발음 정확도가 임계치 미만인 경우, 인공지능 모델(111)의 출력을 기반으로 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0082] 예컨대, 프로세서(120)는 일 사용자의 복수의 텍스트에 대한 음성과 관련하여, 타겟 음성의 변환 없이 의사소통

이 가능한 텍스트에 높은 발음 정확도를 부여하고, 그렇지 않은 텍스트에 낮은 발음 정확도를 부여할 수 있다. 예를 들어, 음성 인식 결과에 따라 획득된 결과 텍스트와 사용자가 발화를 의도한 텍스트 간의 차이가 클수록 발음 정확도가 낮은 것으로 식별될 수 있다. 여기서, 음성 인식은, 일반적인 정상인의 음성을 인식하여 텍스트를 변환하도록 훈련된 적어도 하나의 제2 음성 인식 모델에 의해 수행될 수 있다.

[0083] 그 결과, 일 텍스트에 대해 발음 정확도가 임계치 이상인 경우, 프로세서(120)는 해당 텍스트가 타겟 음성의 변환 없이 의사소통이 가능한 텍스트인 것으로 판단하고, 해당 텍스트에 대한 사용자의 음성을 변환 없이 출력한다.

[0084] 반면, 일 텍스트에 대해 발음 정확도가 임계치 미만인 경우, 프로세서(120)는 해당 텍스트에 대한 사용자의 음성을 타겟 음성으로 변환하여 출력한다.

[0085] 이를 위해, 프로세서(120)는, 복수의 텍스트 각각에 대한 사용자의 음성을 획득하여, 음성 인식을 실시함으로써 복수의 텍스트 각각에 대한 발음 정확도를 매칭하여 저장할 수 있다.

[0086] 한편, 프로세서(120)는, 복수의 텍스트 중, 기 매칭된 발음 정확도가 임계치 미만인 제1 텍스트를 발화하는, 사용자의 새로 입력된 음성의 발음 정확도가 임계치 이상인 경우, 제1 텍스트를 발화하는 사용자의 음성인, 제1 테스트 음성을 일정 횟수만큼 획득하고, 일정 횟수만큼 획득된 제1 테스트 음성의 발음 정확도에 대한 평균이 임계치 이상이면, 제1 테스트 음성의 발음 정확도에 대한 평균을 바탕으로 제1 텍스트에 매칭된 발음 정확도를 업데이트할 수 있다.

[0087] 예를 들어, 사용자의 제1 텍스트에 대한 발음 정확도가 40%이고, 임계치가 50%인 경우를 가정한다.

[0088] 프로세서(120)로 사용자가 제1 텍스트를 발화한 음성이 새로 입력되었을 때, 프로세서(120)가 해당 음성의 제1 텍스트에 대한 발음 정확도가, 임계치보다 높은 60%인 것으로 산출된 경우, 프로세서(120)는 제1 텍스트를 발화하는 사용자의 음성에 있어서, 임계치보다 높아진 발음 정확도의 구체적인 수치를 획득하기 위한 테스트를 실시할 수 있다.

[0089] 구체적으로, 프로세서(120)는 제1 텍스트를 발화하는 사용자의 음성인, 제1 테스트 음성을 일정 횟수(ex. 5회)만큼 획득하고, 5개의 제1 테스트 음성 각각이 제1 텍스트에 대해 갖는 발음 정확도를 산출할 수 있다.

[0090] 이때, 5개의 제1 테스트 음성 각각이 제1 텍스트에 대해 갖는 발음 정확도가 48%, 55%, 52%, 63%, 72%인 경우, 프로세서(120)는 5개의 제1 테스트 음성 각각이 제1 텍스트에 대해 갖는 발음 정확도의 평균 값인 58%를 산출하고, 해당 평균 값을 제1 텍스트에 대해 매칭하여 저장함으로써, 제1 텍스트에 대한 발음 정확도를 업데이트할 수 있다.

[0091] 반면, 프로세서(120)는, 복수의 텍스트 중, 기 매칭된 발음 정확도가 임계치 이상인 제2 텍스트를 발화하는, 사용자의 새로 입력된 음성의 발음 정확도가 임계치 미만인 경우, 제2 텍스트를 발화하는 사용자의 음성인, 제2 테스트 음성을 일정 횟수만큼 획득하고, 일정 횟수만큼 획득된 테스트 음성의 발음 정확도에 대한 평균이 임계치 미만이면, 제2 테스트 음성의 발음 정확도에 대한 평균을 바탕으로 제2 텍스트에 매칭된 발음 정확도를 업데이트할 수 있다.

[0092] 예를 들어, 사용자의 제2 텍스트에 대한 발음 정확도가 60%이고, 임계치가 50%인 경우를 가정한다.

[0093] 프로세서(120)로 사용자가 제2 텍스트를 발화한 음성이 새로 입력되었을 때, 프로세서(120)가 해당 음성의 제2 텍스트에 대한 발음 정확도가, 임계치보다 낮은 32%인 것으로 산출된 경우, 프로세서(120)는 제2 텍스트를 발화하는 사용자의 음성에 있어서, 임계치보다 낮아진 발음 정확도의 구체적인 수치를 획득하기 위한 테스트를 실시할 수 있다.

[0094] 구체적으로, 프로세서(120)는 제2 텍스트를 발화하는 사용자의 음성인, 제2 테스트 음성을 일정 횟수(ex. 5회)만큼 획득하고, 5개의 제2 테스트 음성 각각이 제2 텍스트에 대해 갖는 발음 정확도를 산출할 수 있다.

[0095] 이때, 5개의 제2 테스트 음성 각각이 제2 텍스트에 대해 갖는 발음 정확도가 44%, 55%, 48%, 33%, 28%인 경우, 프로세서(120)는 5개의 제2 테스트 음성 각각이 제2 텍스트에 대해 갖는 발음 정확도의 평균 값인 41.6%를 산출하고, 해당 평균 값을 제2 텍스트에 대해 매칭하여 저장함으로써, 제2 텍스트에 대한 발음 정확도를 업데이트할 수 있다.

[0096] 일 실시예로, 사용자의 음성에 복수의 텍스트가 획득된 경우, 프로세서(120)는 각각의 텍스트에 대한 발음 정확도를 식별하고, 발음 정확도를 바탕으로 각각의 텍스트에 대한 타겟 음성 생성 여부를 판단할 수 있다.

[0097] 구체적으로, 프로세서(120)는, 제2 음성의 발화 대상인 텍스트를 구성하는 제1 서브 텍스트 및 제2 서브 텍스트를 획득하고, 제2 음성의 오디오 데이터 중 제1 서브 텍스트 및 제2 서브 텍스트 각각이 매칭된 오디오 구간을 식별한다.

[0098] 이후, 프로세서(120)는, 제1 서브 텍스트에 대한 제1 오디오 구간의 발음 정확도 및 제2 서브 텍스트에 대한 제2 오디오 구간의 발음 정확도를 식별한다.

[0099] 이때, 제1 서브 텍스트에 대한 발음 정확도가 임계치 미만이고, 제2 서브 텍스트에 대한 발음 정확도가 임계치 이상이면, 인공지능 모델(111)의 출력을 기반으로 제1 서브 텍스트에 대한 타겟 음성을 획득하고, 제1 서브 텍스트에 대한 타겟 음성 및 제2 오디오 구간을 포함하는 오디오 데이터를 바탕으로, 제2 음성이 변환된 음성을 출력할 수 있다.

[0100] 한편, 텍스트는, 도 5, 7, 및 9에 도시된 바와 같이, 단음절(monosyllabic word/language), 단모음(monopthong) 및 단어(word)일 수 있으며, 다양한 실시예로써, 음소(phoneme), 문장(sentence), 형태소(morpheme) 등일 수 있다.

[0101] 한편, 본 발명의 이 면에 따른 딥러닝 기반 구음 장애 음성 개선 변환 시스템을 제어하기 위한 인공지능 모델(111)의 학습에 있어서, 서버(300)는 일 사용자 단말(200)로부터, 일 텍스트를 발화하는 사용자의 제1 음성의 오디오 데이터를 획득하고, 텍스트를 발화하는 정상적인 음성의 오디오 데이터를 획득하여, 제1 음성의 오디오 데이터 및 정상적인 음성의 오디오 데이터를 바탕으로, 타겟 음성을 출력하도록 인공지능 모델(111)을 훈련시킨다.

[0102] 한편, 사용자 단말(200)로부터 사용자의 제2 음성을 수신(S210)하면, 서버(300)는 수신된 제2 음성의 오디오 데이터를 인공지능 모델(111)에 입력하고, 인공지능 모델(111)로부터 제2 음성이 변환된 타겟 음성을 획득(S220)할 수 있다.

[0103] 이때, 서버(300)는 사용자 단말(200)을 통해, 제2 음성이 변환된 타겟 음성을 출력(S230)한다.

[0104] 추가로, 서버(300)가 단계 S220을 수행하는데 있어서, 서버(300)는 제1 음성의 오디오 데이터로부터 추출된, 복수의 제1 특징 벡터와, 제2 음성의 오디오 데이터로부터 추출된 복수의 제2 특징 벡터를 비교하여, 제2 음성의 발화 대상인 텍스트를 식별함으로써, 식별된 텍스트에 매칭되도록, 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0105] 이때, 복수의 특징 벡터는, 일 오디오 데이터로부터 추출된 복수의 음향(aqoustic) 특징 벡터 및 복수의 MFCC(Mel Frequency Cepstral Coefficients) 특징 벡터를 포함할 수 있다.

[0106] 또한, 서버(300)가, 타겟 음성을 출력하도록 인공지능 모델(111)을 훈련시키는데 있어서, 서버(300)는 제1 음성에 대한 오디오 데이터의 주파수 대역을 사용자 표준 데이터로 획득하고, 사용자 표준 데이터의 주파수 대역을 기반으로, 사용자의 제2 음성이 변환된 타겟 음성을 획득할 수 있다.

[0107] 실시예로써, 프로세서는, 사용자의 구음 장애 유형을 식별하고, 구음 장애 유형에 따른 발음하기 어려운 텍스트를 분류하여, 해당 텍스트에 대한 제1 음성을 획득하여 추출한 오디오 데이터를 바탕으로 인공지능 모델(111)을 학습할 수 있다.

[0108] 이를 위해, 본 발명 일 면 및/또는 이 면에 따른 인공지능 모델(111)은, 선택된 특징 벡터들을 적어도 하나의 구음 장애 유형에 따라 분류하여, 구음 장애 유형에 따라 군집화를 실시하는데 있어서, 구음 장애 유형을 분류하기 위한, 별도의 제1 모델 및 제2 모델을 더 포함할 수 있다.

[0109] 구체적으로, 제1 모델은 입력된 복수의 특징 벡터 중 적어도 하나의 특징 벡터를 선택하고, 프로세서(120)는 선택된 특징 벡터를 제2 모델에 입력한다.

[0110] 이때, 제1 모델은 ica(: 독립 성분 분석법, independent component analysis), pca(: 주성분 분석법, principal component analysis), rp(: 랜덤 투영법, random projection), 및 dae(: 딥 오토인코더, deep auto-encoder)를 포함하는 방법 중 적어도 하나의 방법으로 적어도 하나의 특징 벡터를 선택할 수 있다.

[0111] 제2 모델은 입력된 특징 벡터를 분류함으로써 구음 장애 유형을 판단할 수 있으며, svm(: 서포트 벡터 머신, support vector machine), rf(: 무작위 (결정)숲, random forest), mlp(: 다층 퍼셉트론, multi-layer perceptron) 및 cnn(: 합성곱 신경망, convolutional deep neural networks)를 포함하는 분류 방법 중 적어도

하나의 방법으로 특징 벡터를 분류할 수 있다.

[0112] 제1 모델이 선택하는 특징 벡터는, 적어도 하나의 구음 장애 유형에 해당하는 환자의 오디오 데이터가 정상인의 오디오 데이터와 비교된 결과에 따라 설정될 수 있다.

[0113] 예를 들어, 정상인의 오디오 데이터의 특징 벡터 각각과 구음 장애를 가진 환자의 오디오 데이터의 특징 벡터 각각을 비교했을 때, 임계치 이상의 차이를 가지는 특징 벡터가 제1 모델이 선택하는 특징 벡터일 수 있다.

[0114] 구체적인 예로, 복수의 정상인의 오디오 데이터에 대하여 제1 특징 벡터의 평균 값이 산출되고, 복수의 환자의 오디오 데이터에 대하여 제1 특징 벡터의 평균 값이 산출될 수 있으며, 만약 상술한 평균 값들 간의 차이가 일정 수치 이상인 경우, 제1 특징 벡터는 제1 모델에 의해 선택되는 특징 벡터로 설정될 수 있다.

[0115] 한편, 일 실시 예로, 구음 장애 유형에 따라, 서로 다른 그룹의 특징 벡터에 대한 정보가 저장될 수 있다.

[0116] 예를 들어, 제1 구음 장애 유형에 대해서는 제1 그룹의 특징 벡터가 매칭되는 것으로 설정되고, 제2 구음 장애 유형에 대해서는 제2 그룹의 특징 벡터가 매칭되는 것으로 설정될 수 있다.

[0117] 구체적으로, 프로세서(120)는, 구음 장애 유형 별로, 정상인과의 차이가 임계치 이상인 특징 벡터를 식별하여, 해당 특징 벡터를 구음 장애 유형에 매칭되는 그룹으로 저장할 수 있다.

[0118] 여기서, 각 구음 장애 유형에 매칭되는 복수의 (특징 벡터) 그룹에 대해, 제2 모델로 입력되는 순서가 설정될 수 있으며, 순서는, 각 구음 장애 유형의 발생 빈도에 따라 설정될 수 있다.

[0119] 예를 들어, 제1 구음 장애 유형, 제2 구음 장애 유형, 제3 구음 장애 유형의 순서로 발생 빈도가 높은 경우를 가정한다.

[0120] 이 경우, 프로세서(120)는 사용자의 오디오 데이터로부터 먼저 제1 구음 장애 유형에 매칭되는 제1 특징 벡터 그룹을 추출하도록 제1 모델을 제어하고, 추출된 제1 특징 벡터 그룹을 제2 모델로 입력할 수 있다. 그 결과, 사용자가 제1 구음 장애 유형에 해당하는지 여부가 출력될 수 있다.

[0121] 다음으로, 프로세서(120)는 제1 특징 벡터 그룹의 다음 순서인 제2 특징 벡터 그룹을 제2 모델로 입력하여, 그 결과, 사용자가 제2 구음 장애 유형에 해당하는지 여부가 출력될 수 있으며, 프로세서(120)에 의해, 다음 순서인 제3 특징 벡터 그룹이 제2 모델에 입력되어, 그 결과인, 사용자가 제3 구음 장애 유형에 해당하는지 여부가 출력된다.

[0122] 추가 실시예로, 프로세서(120)는 구음 장애 유형 간의 연관성에 따라 각 특징 벡터 그룹이 제2 모델로 입력되는 순서를 유동적으로 변경할 수 있다.

[0123] 구체적으로, 프로세서(120)는 각 구음 장애 유형의 발생 이력을 바탕으로 둘 이상의 구음 장애 유형이 동시에 발생한 빈도/확률을 식별할 수 있다. 그리고, 프로세서(120)는 식별된 빈도/확률에 따라 구음 장애 유형 간의 매칭 여부를 설정할 수 있다.

[0124] 예를 들어, 제1 구음 장애 유형과 제3 구음 장애 유형이 동시에 발생한 빈도가 일정 수 이상인 경우, 프로세서(120)는 제1 구음 장애 유형과 제3 구음 장애 유형이 서로 매칭되는 것으로 설정할 수 있다.

[0125] 관련하여, 일 실시 예로, 제1 특징 벡터 그룹, 제2 특징 벡터 그룹, 제3 특징 벡터 그룹의 순서대로 제2 모델로 입력되는 것으로 기설정된 상황을 가정한다.

[0126] 이 경우, 일반적으로, 프로세서(120)는 제1 특징 벡터 그룹을 제2 모델로 입력하여, 사용자가 제1 구음 장애 유형에 해당하는지 여부를 출력하고, 이어서, 제1 특징 벡터 그룹의 다음 순서인 제2 특징 벡터 그룹을 제2 모델로 입력하여, 사용자가 제2 구음 장애 유형에 해당하는지 여부를 출력한다.

[0127] 다만, 제1 특징 벡터 그룹이 입력된 결과 사용자가 제1 구음 장애 유형으로 판단되면, 프로세서(120)는, 제1 특징 벡터 그룹의 다음 순번인 제2 특징 벡터 그룹이 아닌, 제3 특징 벡터 그룹을 제2 모델로 입력할 수 있다.

[0128] 구체적으로, 프로세서(120)는, 사용자가 제1 구음 장애 유형으로 판단된 경우, 제1 구음 장애 유형에 매칭된 구음 장애 유형인 제3 구음 장애 유형을 식별하고, 제3 구음 장애 유형에 매칭된 제3 특징 벡터 그룹을 제2 모델로 입력한다.

[0129] 그 결과, 사용자가 제3 구음 장애 유형에 해당하는지 여부를 출력할 수 있다.

[0130] 이후, 사용자가 제3 구음 장애 유형이 아닌 것으로 판단되거나, 또는 사용자가 제3 구음 장애 유형으로 판단되

되, 제3 구음 장애 유형에 매칭된 구음 장애 유형이 존재하지 않는 경우, 프로세서(120)는 제1 특징 벡터 그룹의 다음 순번인 제2 특징 벡터 그룹을 제2 모델로 입력하여, 사용자가 제2 구음 장애 유형에 해당하는지 여부를 출력할 수 있다.

[0131] 반면, 사용자가 제3 구음 장애 유형으로 판단되되, 제3 구음 장애 유형에 매칭된 구음 장애 유형으로 제4 구음 장애 유형이 식별된 경우, 프로세서(120)는, 제4 구음 장애 유형에 매칭된 제4 특징 벡터 그룹을 제2 모델에 우선 입력할 수 있다.

[0132] 추가로, 제1 모델은 오디오 데이터로부터 추출된, 복수의 음향 특징 벡터 및 복수의 MFCC 특징 벡터 중, 텍스트에 따라 적어도 하나의 특징 벡터를 선택할 수 있다.

[0133] 구체적으로, 제1 모델은, 텍스트에 따라 적어도 하나의 특징 벡터를 선택하고, 제2 모델은, 선택된 특징 벡터를 바탕으로, 텍스트를 발화하는 정상인의 음성의 오디오 데이터를 오디오 데이터와 비교한다.

[0134] 이를 위해, 오디오 데이터는 텍스트에 매칭되어 프로세서(120)에 입력되며, 프로세서(120)는 오디오 데이터를 전처리하여 복수의 특징 벡터를 텍스트와 함께 제1 모델로 입력한다.

[0135] 한편, 프로세서(120)는, 사용자의 실제 구음 장애 유형이 인공지능 모델(111)로부터 출력된 구음 장애 유형과 일치하지 않는 경우, 텍스트에 매칭된 오차 발생 횟수를 업데이트하고, 오차 발생 횟수가 일정 값을 초과하면, 텍스트에 대하여 다른 특징 벡터를 선택하도록 제1 모델을 업데이트할 수 있다.

[0136] 이때, 업데이트된 제1 모델을 통한 선택 및 기존 제2 모델에 의한 분류로 획득된 구음 장애 유형이 실게 구음 장애 유형과 일치하지 않는 경우, 프로세서(120)는, 텍스트에 대하여 다른 방식으로 분류를 실시하도록 제2 모델을 업데이트할 수 있다.

[0137] 예를 들어, 기존 구음 장애 분류 방법으로 본 발명에 의한 제1 모델이 ica를 적용하여 98개의 특징 벡터 중 40개의 특징 벡터(1차 선택 특징 벡터)를 선택하고, 제2 모델이 mlp(1차 분류 방식)를 적용하여 분류를 실시하는 경우, 텍스트 "낮새"에 대한 사용자의 구음 장애 유형이 파킨슨인 반면, 제2 모델에 의해 후두로 출력되면, 프로세서(120)는 텍스트 "낮새"에 매칭된 98개의 특징 벡터 중 1차 선택 특징 벡터를 제외한 나머지 58개의 특징 벡터로부터 신규 40개의 특징 벡터(2차 선택 특징 벡터)를 선택하도록 제1 모델을 업데이트 할 수 있다.

[0138] 이 경우, 인공지능 모델(111)은, 업데이트 이후 획득된, 텍스트 "낮새"에 대한 오디오 데이터에 매칭된 복수의 특징 벡터가 입력되면, 기 매칭된 2차 선택 특징 벡터를 자동으로 선택하여 분류를 실시함으로써 텍스트 "낮새"에 대한 구음 장애 분류의 정확도를 높일 수 있다.

[0139] 이때, 텍스트 "낮새"에 대한 제1 모델의 업데이트에도 불구하고 제2 모델로부터의 출력이 파킨슨이 아닌 경우, 프로세서(120)는 제2 모델의 분류 방식을 1차 분류 방식 외의 다른 방식으로 선택할 수 있으며, 다른 방식을 각각 적용하여 후두형으로 분류되는 방식(2차 분류 방식)을 획득하고, 텍스트 "낮새"에 매칭된 제2 모델의 분류 방식을 2차 분류 방식으로 업데이트함으로써, 텍스트 "낮새"에 대한 구음 장애 분류의 정확도를 높일 수 있다.

[0140] 또한, 프로세서(120)는 텍스트에 대한 오류 발생 횟수가 일정 값을 초과하면, 일 텍스트에 대하여 선택되는 특징 벡터의 수를 변동하도록 제1 모델을 업데이트 할 수 있다.

[0141] 예를 들어, 제1 모델 및 제2 모델 각각의 업데이트 이후에도 텍스트 "낮새"에 대한 제2 모델의 구음 장애 분류 결과가 사용자의 구음 장애 유형과 다른 유형인 경우, 총 3회의 오류가 발생되었으므로, 프로세서(120)는 텍스트 "낮새"에 매칭된 오류 발생 횟수를 3회로 업데이트 할 수 있다.

[0142] 이때, 오류 발생 횟수가 2회를 초과한 것에 대해, 프로세서(120)는 일정 값(2회)을 초과한 것으로 판단하고, 텍스트 "낮새"를 발화하는 오디오 데이터에 대해 제1 모델의 특징 벡터 선택 수를 20개로 변동하도록 업데이트 할 수 있다.

[0143] 한편, 본 발명은 하드웨어인 컴퓨터와 결합되어, 본 발명의 이 면에 따른 방법을 수행할 수 있도록 컴퓨터에서 독출가능한 기록매체에 저장된 컴퓨터 프로그램을 포함할 수 있다.

[0144] 도 8은 본 발명의 일 실시예에 다른 서버 구성도이다.

[0145] 도시된 바와 같이, 서버(300)는 메모리(310), 통신부(320) 및 프로세서(330)를 포함할 수 있다.

[0146] 메모리(310)는 서버(300)의 동작에 필요한 각종 프로그램 및 데이터를 저장할 수 있다. 메모리(310)는 비휘발성 메모리(310), 휘발성 메모리(310), 플래시메모리(310)(flash-memory), 하드디스크 드라이브(HDD) 또는 솔리드

스테이트 드라이브(SSD) 등으로 구현될 수 있다.

[0147] 통신부(320)는 외부 장치와 통신을 수행할 수 있다. 특히, 통신부(320)는 와이파이 칩, 블루투스 칩, 무선 통신 칩, NFC칩, 저전력 블루투스 침(BLE 칩) 등과 같은 다양한 통신 칩을 포함할 수 있다. 이때, 와이파이 칩, 블루투스 칩, NFC 칩은 각각 LAN 방식, WiFi 방식, 블루투스 방식, NFC 방식으로 통신을 수행한다. 와이파이 칩이나 블루투스칩을 이용하는 경우에는 SSID 및 세션 키 등과 같은 각종 연결 정보를 먼저 송수신 하여, 이를 이용하여 통신 연결한 후 각종 정보들을 송수신할 수 있다. 무선 통신칩은 IEEE, 지그비, 3G(3rd Generation), 3GPP(3rd Generation Partnership Project), LTE(Long Term Evolution), 5G(5th Generation) 등과 같은 다양한 통신 규격에 따라 통신을 수행하는 칩을 의미한다.

[0148] 프로세서(330)는 메모리(310)에 저장된 각종 프로그램을 이용하여 서버(300)의 전반적인 동작을 제어할 수 있다. 프로세서(330)는 RAM, ROM, 그래픽 처리부, 메인 CPU, 제1 내지 n 인터페이스 및 버스로 구성될 수 있다. 이때, RAM, ROM, 그래픽 처리부, 메인 CPU, 제1 내지 n 인터페이스 등은 버스를 통해 서로 연결될 수 있다.

[0149] RAM은 O/S 및 어플리케이션 프로그램을 저장한다. 구체적으로, 서버(300)가 부팅되면 O/S가 RAM에 저장되고, 사용자가 선택한 각종 어플리케이션 데이터가 RAM에 저장될 수 있다.

[0150] ROM에는 시스템 부팅을 위한 명령어 세트 등이 저장된다. 턴 온 명령이 입력되어 전원이 공급되면, 메인 CPU는 ROM에 저장된 명령어에 따라 메모리(310)에 저장된 O/S를 RAM에 복사하고, O/S를 실행시켜 시스템을 부팅시킨다. 부팅이 완료되면, 메인 CPU는 메모리(310)에 저장된 각종 어플리케이션 프로그램을 RAM에 복사하고, RAM에 복사된 어플리케이션 프로그램을 실행시켜 각종 동작을 수행한다.

[0151] 그래픽 처리부는 연산부(미도시) 및 렌더링부(미도시)를 이용하여 아이템, 이미지, 텍스트 등과 같은 다양한 객체를 포함하는 화면을 생성한다. 여기서, 연산부는 입력부로부터 수신된 제어 명령을 이용하여 화면의 레이아웃에 따라 각 객체들이 표시될 좌표값, 형태, 크기, 컬러 등과 같은 속성값을 연산하는 구성일 수 있다. 그리고, 렌더링부는 연산부에서 연산한 속성값에 기초하여 객체를 포함하는 다양한 레이아웃의 화면을 생성하는 구성이일 수 있다. 이러한 렌더링부에서 생성된 화면은 디스플레이의 디스플레이 영역 내에 표시될 수 있다.

[0152] 메인 CPU는 메모리(310)에 액세스하여, 메모리(310)에 저장된 OS를 이용하여 부팅을 수행한다. 그리고, 메인 CPU는 메모리(310)에 저장된 각종 프로그램, 컨텐츠, 데이터 등을 이용하여 다양한 동작을 수행한다.

[0153] 제1 내지 n 인터페이스는 상술한 각종 구성요소들과 연결된다. 제1 내지 n 인터페이스 중 하나는 네트워크를 통해 외부 장치와 연결되는 네트워크 인터페이스가 될 수도 있다.

[0154] 한편, 나아가, 프로세서(330)는 인공지능 모델을 제어할 수 있다. 이 경우, 프로세서(330)는 인공지능 모델을 제어하기 위한 그래픽 전용 프로세서(예: GPU)를 포함할 수 있음은 물론이다.

[0155] 한편, 본 발명에 따른 인공지능 모델은 교사 지도학습(supervised learning) 또는 비교사 지도학습(unsupervised learning)기반의 모델일 수 있다. 나아가, 본 발명에 따른 인공지능 모델은 SVM(support vector machine), Decision tree, neural network 등 및 이들이 응용된 방법론을 포함할 수 있다.

[0156] 일 실시예로, 본 발명에 따른 인공지능 모델은 학습데이터를 입력하여 학습된 합성곱 신경망(Convolutional deep Neural Networks, CNN) 기반의 인공지능 모델일 수 있다. 다만, 이에 한정되는 것은 아니며, 다양한 인공지능 모델이 본 발명에 적용될 수 있음은 물론이다. 예컨대, DNN(Deep Neural Network), RNN(Recurrent Neural Network), BRDNN(Bidirectional Recurrent Deep Neural Network)과 같은 모델이 인공지능 모델로서 사용될 수 있으나, 이에 한정되지 않는다.

[0157] 이때, 합성곱 신경망(Convolutional deep Neural Networks, CNN)은 최소한의 전처리(preprocess)를 사용하도록 설계된 다계층 퍼셉트론(multilayer perceptrons)의 한 종류이다. 합성곱 신경망은 하나 또는 여러개의 합성곱 계층(convolutional layer)과 그 위에 올려진 일반적인 인공신경망 계층들로 이루어져 있으며, 가중치와 통합 계층(pooling layer)들을 추가로 활용한다. 이러한 구조 덕분에 합성곱 신경망은 2차원 구조의 입력 데이터를 충분히 활용할 수 있다. 또한, 합성곱 신경망은 표준 역전달을 통해 훈련될 수 있다. 합성곱 신경망은 다른 피드포워드 인공신경망 기법들보다 쉽게 훈련되는 편이고 적은 수의 매개변수를 사용한다는 이점이 있다.

[0158] 또한, 심층 신경망(Deep Neural Networks, DNN)은 입력 계층(input layer)과 출력 계층(output layer) 사이에 복수개의 은닉 계층(hidden layer)들로 이뤄진 인공신경망(Artificial Neural Network, ANN)이다.

[0159] 이때, 심층 신경망의 구조는 퍼셉트론(perceptron)으로 구성될 수 있다. 퍼셉트론은 여러 개의 입력 값(input)

과 하나의 프로세서(prosessor), 하나의 출력 값으로 구성된다. 프로세서는 여러 개의 입력 값에 각각 가중치를 곱한 후, 가중치가 곱해진 입력 값들을 모두 합한다. 그 다음 프로세서는 합해진 값을 활성화함수에 대입하여 하나의 출력 값을 출력한다. 만약 활성화함수의 출력 값으로 특정한 값이 나오기를 원하는 경우, 각 입력 값에 곱해지는 가중치를 수정하고, 수정된 가중치를 이용하여 출력 값을 다시 계산할 수 있다. 이때, 각각의 퍼셉트론은 서로 다른 활성화함수를 사용할 수 있다. 또한 각각의 퍼셉트론은 이전 계층에서 전달된 출력들을 입력으로 받아들인 다음, 활성화 함수를 이용해서 출력을 구한다. 구해진 출력은 다음 계층의 입력으로 전달된다. 상술한 바와 같은 과정을 거치면 최종적으로 몇 개의 출력 값을 얻을 수 있다.

[0160] 순환 신경망(Reccurent Neural Network, RNN)은 인공신경망을 구성하는 유닛 사이의 연결이 Directed cycle을 구성하는 신경망을 말한다. 순환 신경망은 앞먹임 신경망과 달리, 임의의 입력을 처리하기 위해 신경망 내부의 메모리를 활용할 수 있다.

[0161] 심층 신뢰 신경망(Deep Belief Networks, DBN)이란 기계학습에서 사용되는 그래프 생성 모형(generative graphical model)으로, 딥 러닝에서는 잠재변수(latent variable)의 다중계층으로 이루어진 심층 신경망을 의미한다. 계층 간에는 연결이 있지만 계층 내의 유닛 간에는 연결이 없다는 특징이 있다.

[0162] 심층 신뢰 신경망은 생성 모형이라는 특성상 선행학습에 사용될 수 있고, 선행학습을 통해 초기 가중치를 학습한 후 역전파 혹은 다른 판별 알고리즘을 통해 가중치의 미조정을 할 수 있다. 이러한 특성은 훈련용 데이터가 적을 때 굉장히 유용한데, 이는 훈련용 데이터가 적을수록 가중치의 초기값이 결과적인 모델에 끼치는 영향이 세지기 때문이다. 선행학습된 가중치 초기값은 임의로 설정된 가중치 초기값에 비해 최적의 가중치에 가깝게 되고 이는 미조정 단계의 성능과 속도향상을 가능케 한다.

[0163] 상술한 인공지능 및 그 학습방법에 관한 내용은 예시를 위하여 서술된 것이며, 상술한 실시 예들에서 이용되는 인공지능 및 그 학습방법은 제한되지 않는다. 예를 들어, 당 업계의 통상의 기술자가 동일한 과제해결을 위하여 적용할 수 있는 모든 종류의 인공지능 기술 및 그 학습방법이 개시된 실시 예에 따른 시스템을 구현하는 데 활용될 수 있다.

[0164] 한편, 프로세서(330)는 하나 이상의 코어(core, 미도시) 및 그래픽 처리부(미도시) 및/또는 다른 구성 요소와 신호를 송수신하는 연결 통로(예를 들어, 버스(bus) 등)를 포함할 수 있다.

[0165] 일 실시예에 따른 프로세서(330)는 메모리(310)에 저장된 하나 이상의 인스트럭션을 실행함으로써, 본 발명과 관련하여 설명된 방법을 수행한다.

[0166] 예를 들어, 프로세서(330)는 메모리(310)에 저장된 하나 이상의 인스트럭션을 실행함으로써 신규 학습용 데이터를 획득하고, 학습된 모델을 이용하여, 상기 획득된 신규 학습용 데이터에 대한 테스트를 수행하고, 상기 테스트 결과, 라벨링된 정보가 소정의 제1 기준값 이상의 정확도로 획득되는 제1 학습용 데이터를 추출하고, 상기 추출된 제1 학습용 데이터를 상기 신규 학습용 데이터로부터 삭제하고, 상기 추출된 학습용 데이터가 삭제된 상기 신규 학습용 데이터를 이용하여 상기 학습된 모델을 다시 학습시킬 수 있다.

[0167] 한편, 프로세서(330)는 프로세서(330) 내부에서 처리되는 신호(또는, 데이터)를 일시적 및/또는 영구적으로 저장하는 램(RAM: Random Access Memory, 미도시) 및 롬(ROM: Read-Only Memory, 미도시)을 더 포함할 수 있다. 또한, 프로세서(330)는 그래픽 처리부, 램 및 롬 중 적어도 하나를 포함하는 시스템온칩(SoC: system on chip) 형태로 구현될 수 있다.

[0168] 메모리(310)에는 프로세서(330)의 처리 및 제어를 위한 프로그램들(하나 이상의 인스트럭션들)을 저장할 수 있다. 메모리(310)에 저장된 프로그램들은 기능에 따라 복수 개의 모듈들로 구분될 수 있다.

[0169] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

[0170] 본 발명의 구성 요소들은 하드웨어인 컴퓨터와 결합되어 실행되기 위해 프로그램(또는 애플리케이션)으로 구현되어 매체에 저장될 수 있다. 본 발명의 구성 요소들은 소프트웨어 프로그래밍 또는 소프트웨어 요소들로 실행될 수 있으며, 이와 유사하게, 실시 예는 데이터 구조, 프로세스들, 루틴들 또는 다른 프로그래밍 구성들의 조

합으로 구현되는 다양한 알고리즘을 포함하여, C, C++, 자바(Java), 어셈블러(assembler) 등과 같은 프로그래밍 또는 스크립팅 언어로 구현될 수 있다. 기능적인 측면들은 하나 이상의 프로세서들에서 실행되는 알고리즘으로 구현될 수 있다.

[0171] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

부호의 설명

[0172] 100 : 전자 장치

110 : 메모리

111 : 인공지능 모델

120 : 프로세서

130 : 음성 입출력부

130a : 마이크

130b : 스피커

140 : 디스플레이

150 : 통신부

200 : 사용자 단말

300 : 서버

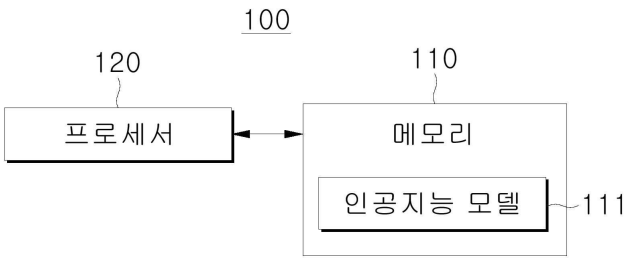
310 : 메모리

320 : 통신부

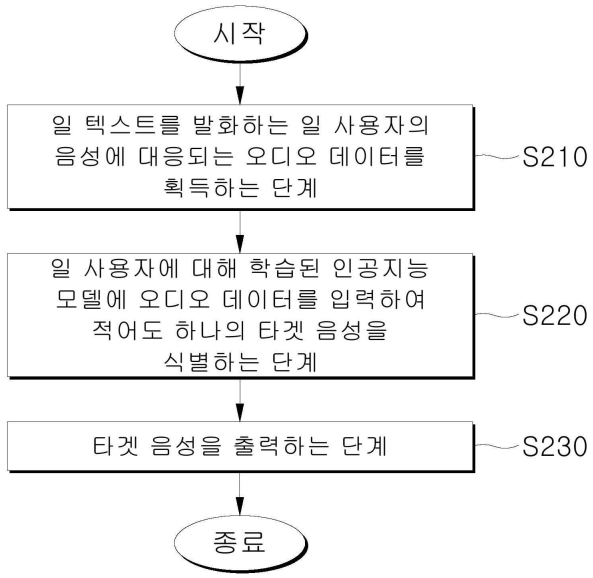
330 : 프로세서

도면

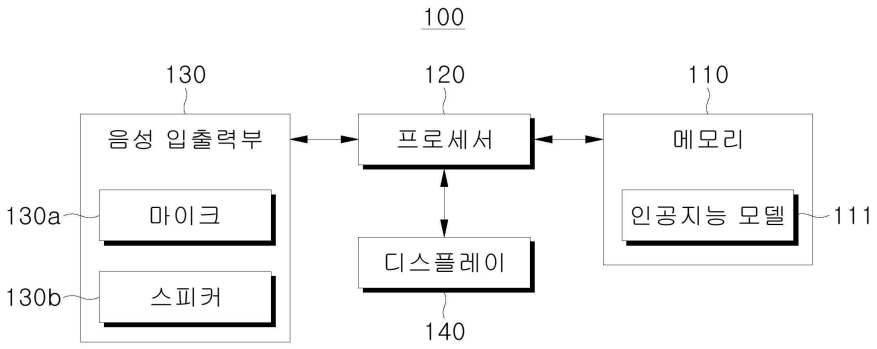
도면1



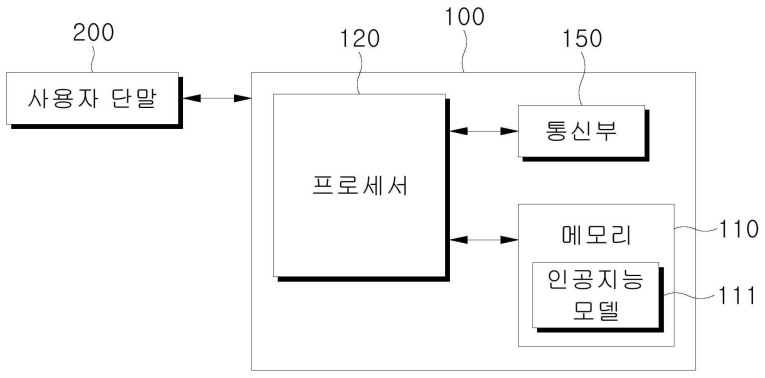
도면2


시작
일 텍스트를 발화하는 일 사용자의 음성에 대응되는 오디오 데이터를 획득하는 단계
S210
일 사용자에 대해 학습된 인공지능 모델에 오디오 데이터를 입력하여 적어도 하나의 타겟 음성을 식별하는 단계
S220
타겟 음성을 출력하는 단계
S230
종료

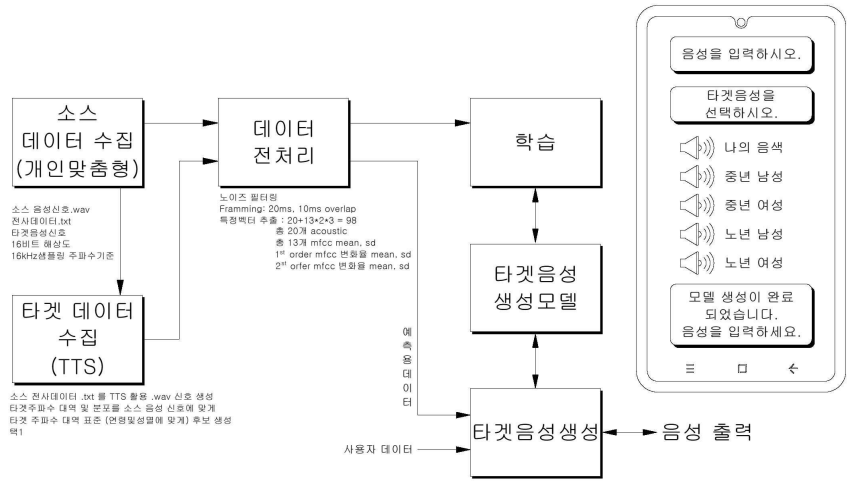
도면3


100
130
120
110
음성 입출력부
프로세서
메모리
130a
마이크
인공지능 모델
111
130b
스피커
디스플레이
140

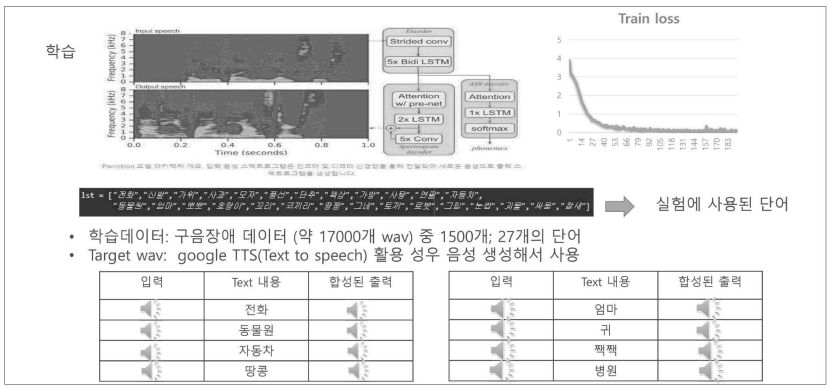
도면4


200
120
100
150
사용자 단말
프로세서
통신부
메모리
110
인공지능 모델
111

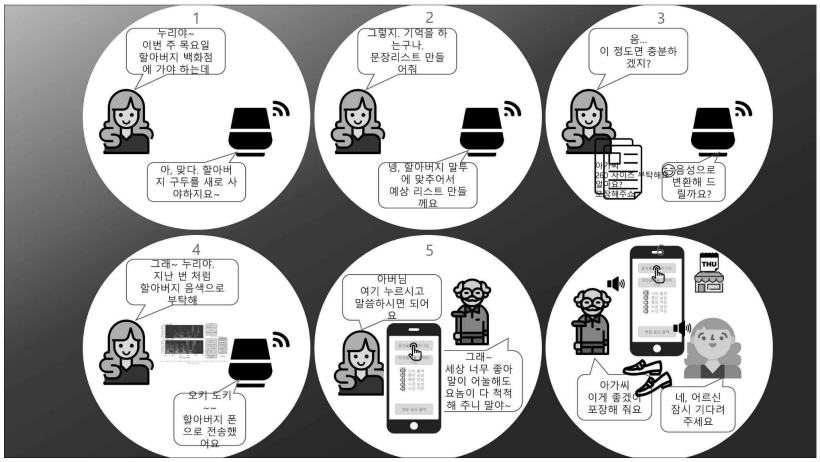
도면5


소스
데이터 수집
(개인맞춤형)
데이터
전처리
학습
타겟음성
생성모델
타겟 데이터
수집
(TTS)
타겟음성생성
음성 출력
사용자 데이터
음성을 입력하시오.
타겟음성을
선택하시오.
나의 음색
중년 남성
중년 여성
노년 남성
노년 여성
모델 생성이 완료
되었습니다.
음성을 입력하세요.

도면6


학습
Train loss
실험에 사용된 단어
학습데이터: 구음장애 데이터 (약 17000개 wav) 중 1500개; 27개의 단어
Target wav: google TTS(Text to speech) 활용 성우 음성 생성해서 사용
입력
Text 내용
합성된 출력
전화
동물원
자동차
땅콩
엄마
귀
짝짝
병원

도면7


누리야~
이번 주 목요일
할아버지 백화점
에 가야 하는데
아, 맞다. 할아버
지 구두를 새로 사
야하지요~
그렇지. 기억을 하
는구나.
문장리스트 만들
어줘
넹, 할아버지 말투
에 맞추어서
예상 리스트 만들
께요
음...
이 정도면 충분하
겠지?
음성으로
변환해 드
릴까요?
그래~ 누리야.
지난 번 처럼
할아버지 음색으로
부탁해
오키 도키
~~
할아버지 폰
으로 전송했
어요
아버님
여기 누르시고
말씀하시면 되어
요
그래~
세상 너무 좋아
말이 어눌해도
요놈이 다 척척
해 주니 말야~
아기씨
이게 중지~
포장해 줘요
네, 어르신
잠시 기다려
주세요

도면8

