(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0027331
(43) 공개일자 2020년03월12일

(51) 국제특허분류(Int. Cl.)
G10L 13/08 (2006.01) *G06F 40/40* (2020.01)
G10L 15/00 (2006.01) *G10L 15/04* (2006.01)
(52) CPC특허분류
G10L 13/08 (2013.01)
G06F 40/40 (2020.01)
(21) 출원번호 10-2018-0105487
(22) 출원일자 2018년09월04일
심사청구일자 2018년09월04일

(71) 출원인
엘지전자 주식회사
서울특별시 영등포구 여의대로 128 (여의도동)
(72) 발명자
채종훈
서울특별시 서초구 양재대로11길 19 LG전자 특허센터
박용철
서울특별시 서초구 양재대로11길 19 LG전자 특허센터
(뒷면에 계속)
(74) 대리인
허용록

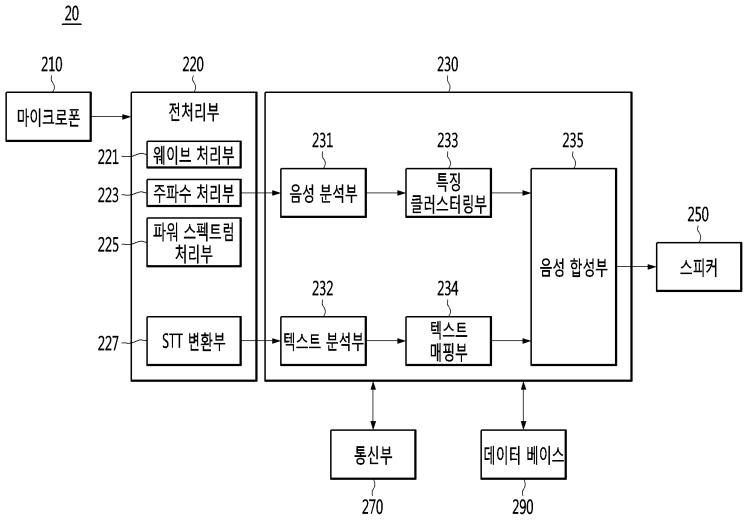
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **음성 합성 장치**

(57) 요 약

본 발명의 일 실시 예에 따른 음성 합성 장치는 음성 및 상기 음성에 대응하는 텍스트를 저장하는 데이터 베이스 및 상기 데이터 베이스에 저장된 제1 언어의 음성의 음색 및 특성 정보를 추출하고, 상기 추출된 특성 정보에 기초하여, 화자의 발화 유형을 분류하고, 상기 음색 및 상기 발화 유형을 포함하는 화자 분석 정보를 생성하고, 상기 제1 언어의 음성에 대응하는 텍스트를 제2 언어로 번역하고, 상기 화자 분석 정보를 이용하여, 상기 제2 언어로 번역된 텍스트를 상기 제2 언어의 음성으로 합성하는 프로세서를 포함한다.

대 표 도 - 도2



(52) CPC특허분류
G10L 15/005 (2013.01)
G10L 15/04 (2013.01)
G10L 19/02 (2013.01)

(72) 발명자
양시영
서울특별시 서초구 양재대로11길 19 LG전자 특허센터
장주영
서울특별시 서초구 양재대로11길 19 LG전자 특허센터
한성민
서울특별시 서초구 양재대로11길 19 LG전자 특허센터

명 세 서

청구범위

청구항 1

음성 합성 장치에 있어서,

음성 및 상기 음성에 대응하는 텍스트를 저장하는 데이터 베이스; 및

상기 데이터 베이스에 저장된 제1 언어의 음성의 음색 및 특성 정보를 추출하고, 상기 추출된 특성 정보에 기초하여, 화자의 발화 유형을 분류하고, 상기 음색 및 상기 발화 유형을 포함하는 화자 분석 정보를 생성하고, 상기 제1 언어의 음성에 대응하는 텍스트를 제2 언어로 번역하고, 상기 화자 분석 정보를 이용하여, 상기 제2 언어로 번역된 텍스트를 상기 제2 언어의 음성으로 합성하는 프로세서를 포함하는

음성 합성 장치.

청구항 2

제1항에 있어서,

상기 제1 언어의 음성의 파형, 상기 제1 언어의 음성의 주파수 대역 및 상기 제1 언어의 음성의 파워 스펙트럼을 추출하는 전처리부를 더 포함하는

음성 합성 장치.

청구항 3

제2항에 있어서,

상기 프로세서는

상기 전처리부에서 추출된 상기 음성의 파형, 상기 음성의 주파수 대역 및 상기 파워 스펙트럼 중 하나 이상을 이용하여, 상기 음색 및 상기 특성 정보를 추출하는

음성 합성 장치.

청구항 4

제3항에 있어서,

상기 특성 정보는

화자의 성별 정보, 음의 높낮이, 화자의 발화 속도, 화자의 감정을 포함하는

음성 합성 장치.

청구항 5

제4항에 있어서,

상기 프로세서는

상기 특성 정보를 구성하는 복수의 유형 항목들 각각에 가중치를 두고, 상기 가중치가 반영된 상기 발화 유형을 생성하는

음성 합성 장치.

청구항 6

제5항에 있어서,

상기 프로세서는

상기 제2 언어로 번역된 텍스트에 상기 음색을 반영한 번역 합성 음성을 생성하고,

상기 번역 합성 음성에 상기 생성된 발화 유형을 적용하여, 최종 합성 음성을 생성하는

음성 합성 장치.

청구항 7

제1항에 있어서,

상기 프로세서는

상기 제1 언어의 음성에 대응하는 텍스트에 복수의 문장들이 포함된 경우, 복수의 문장들 각각으로부터 상기 특성 정보를 추출하고, 추출된 특성 정보에 기초하여, 상기 복수의 문장들 각각에 대응하는 복수의 발화 유형들을 생성하는

음성 합성 장치.

청구항 8

제7항에 있어서,

상기 프로세서는

상기 복수의 문장들이 상기 제2 언어로 번역된 복수의 번역 문장들 각각에 대응하는 발화 유형을 적용하여, 상기 제2 언어로 합성된 음성을 생성하는

음성 합성 장치.

청구항 9

음성 합성 장치의 동작 방법에 있어서,

데이터 베이스에 저장된 제1 언어의 음성의 음색 및 특성 정보를 추출하는 단계;

상기 추출된 특성 정보에 기초하여, 화자의 발화 유형을 분류하는 단계;

상기 음색 및 상기 발화 유형을 포함하는 화자 분석 정보를 생성하는 단계;

상기 제1 언어의 음성에 대응하는 텍스트를 제2 언어로 번역하는 단계; 및

상기 화자 분석 정보를 이용하여, 상기 제2 언어로 번역된 텍스트를 상기 제2 언어의 음성으로 합성하는 단계를 포함하는

음성 합성 장치의 동작 방법.

청구항 10

제9항에 있어서,

상기 제1 언어의 음성의 파형, 상기 제1 언어의 음성의 주파수 대역 및 상기 제1 언어의 음성의 파워 스펙트럼을 추출하는 단계를 더 포함하는

음성 합성 장치의 동작 방법.

청구항 11

제10항에 있어서,

상기 특성 정보를 추출하는 단계는

상기 추출된 상기 음성의 파형, 상기 음성의 주파수 대역 및 상기 파워 스펙트럼 중 하나 이상을 이용하여, 상기 음색 및 상기 특성 정보를 추출하는 단계를 포함하고,

상기 특성 정보는

화자의 성별 정보, 음의 높낮이, 화자의 발화 속도, 화자의 감정을 포함하는

음성 합성 장치의 동작 방법.

청구항 12

제11항에 있어서,

상기 발화 유형을 분류하는 단계는

상기 상기 특성 정보를 구성하는 복수의 유형 항목들 각각에 가중치를 부여하는 단계; 및

상기 가중치가 반영된 상기 발화 유형을 생성하는 단계를 포함하는

음성 합성 장치의 동작 방법.

청구항 13

제12항에 있어서,

상기 제2 언어의 음성으로 합성하는 단계는

상기 제2 언어로 번역된 텍스트에 상기 음색을 반영한 번역 합성 음성을 생성하는 단계와

상기 번역 합성 음성에 상기 생성된 발화 유형을 적용하여, 최종 합성 음성을 생성하는 단계를 포함하는

음성 합성 장치의 동작 방법.

청구항 14

제9항에 있어서,

상기 특성 정보를 추출하는 단계는

상기 제1 언어의 음성에 대응하는 텍스트에 복수의 문장들이 포함된 경우, 복수의 문장들 각각으로부터 상기 특성 정보를 추출하는 단계를 포함하고,

상기 발화 유형을 분류하는 단계는

추출된 특성 정보에 기초하여, 상기 복수의 문장들 각각에 대응하는 복수의 발화 유형들을 생성하는 단계를 포함하는

음성 합성 장치의 동작 방법.

청구항 15

제14항에 있어서,

상기 제2 언어의 음성으로 합성하는 단계는

상기 복수의 문장들이 상기 제2 언어로 번역된 복수의 번역 문장들 각각에 대응하는 발화 유형을 적용하여, 상기 제2 언어로 합성된 음성을 생성하는 단계를 포함하는

음성 합성 장치의 동작 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 음성 합성 장치에 관한 것이다.

배 경 기 술

[0002] 인공 지능(artificial intelligence)은 인간의 지능으로 할 수 있는 사고, 학습, 자기계발 등을 컴퓨터가 할 수 있도록 하는 방법을 연구하는 컴퓨터 공학 및 정보기술의 한 분야로, 컴퓨터가 인간의 지능적인 행동을 모방할

수 있도록 하는 것을 의미한다.

[0003] 또한, 인공지능은 그 자체로 존재하는 것이 아니라, 컴퓨터 과학의 다른 분야와 직간접으로 많은 관련을 맺고 있다. 특히 현대에는 정보기술의 여러 분야에서 인공지능적 요소를 도입하여, 그 분야의 문제 풀이에 활용하려는 시도가 매우 활발하게 이루어지고 있다.

[0004] 인공 지능을 이용한 음성 에이전트 서비스는 사용자의 음성에 응답하여, 사용자에게 필요한 정보를 제공하는 서비스이다.

[0005] 그러나, 종래의 외국어를 더빙 또는 통역하여, 한국어로 제공해 주는 음성 에이전트 서비스는 원 발화자의 목소리나 스타일과 관계 없는 특정 성우가 발화하는 일괄적인 음성을 제공하는데 그쳤다.

[0006] 이에 따라, 영화와 같은 컨텐트의 제공 시, 시청자는 원 발화자의 발화 느낌을 그대로, 전달받을 수 없는 한계가 있었다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 문제 및 다른 문제를 해결하는 것을 목적으로 한다.

[0008] 본 발명은 원 발화자의 목소리나 스타일을 반영하여, 더빙이나, 통역 음성을 시청자에게 제공할 수 있는 음성 합성 장치의 제공을 목적으로 한다.

과제의 해결 수단

[0009] 본 발명의 일 실시 예에 따른 음성 합성 장치는 음성 및 상기 음성에 대응하는 텍스트를 저장하는 데이터 베이스 및 상기 데이터 베이스에 저장된 제1 언어의 음성의 음색 및 특성 정보를 추출하고, 상기 추출된 특성 정보에 기초하여, 화자의 발화 유형을 분류하고, 상기 음색 및 상기 발화 유형을 포함하는 화자 분석 정보를 생성하고, 상기 제1 언어의 음성에 대응하는 텍스트를 제2 언어로 번역하고, 상기 화자 분석 정보를 이용하여, 상기 제2 언어로 번역된 텍스트를 상기 제2 언어의 음성으로 합성하는 프로세서를 포함한다.

[0010] 본 발명의 일 실시 예에 따른 음성 합성 장치의 동작 방법은 데이터 베이스에 저장된 제1 언어의 음성의 음색 및 특성 정보를 추출하는 단계와 상기 추출된 특성 정보에 기초하여, 화자의 발화 유형을 분류하는 단계와 상기 음색 및 상기 발화 유형을 포함하는 화자 분석 정보를 생성하는 단계와 상기 제1 언어의 음성에 대응하는 텍스트를 제2 언어로 번역하는 단계 및 상기 화자 분석 정보를 이용하여, 상기 제2 언어로 번역된 텍스트를 상기 제2 언어의 음성으로 합성하는 단계를 포함한다.

[0011] 본 발명의 적용 가능성의 추가적인 범위는 이하의 상세한 설명으로부터 명백해질 것이다. 그러나 본 발명의 사상 및 범위 내에서 다양한 변경 및 수정은 당업자에게 명확하게 이해될 수 있으므로, 상세한 설명 및 본 발명의 바람직한 실시 예와 같은 특정 실시 예는 단지 예시로 주어진 것으로 이해되어야 한다.

발명의 효과

[0012] 본 발명의 실시 예에 따르면, 외국어로 발화하는 컨텐트의 시청 시, 원 발화자의 목소리 및 스타일이 반영된 더빙이나, 통역이 제공되어, 시청자에게 향상된 음성 서비스 경험을 줄 수 있다.

[0013] 또한, 본 발명의 실시 예에 따르면, 실제 화자가 번역된 음성으로 발화하는 것과 같은 효과가 주어져, 시청자의 컨텐트 시청 몰입도가 향상될 수 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시 예에 따른 음성 합성 시스템의 구성을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시 예에 따른 음성 합성 장치의 구성을 설명하기 위한 블록도이다.

도 3은 본 발명의 일 실시 예에 따른 음성 합성 장치의 동작 방법을 설명하기 위한 흐름도이다.

도 4는 본 발명의 일 실시 예에 따라, 화자의 제1 언어로된 음성으로부터 음성의 특성 정보를 추출하는 예를 설명하는 도면이다.

도 5a 및 도 5b는 본 발명의 일 실시 예에 따라, 음성의 특성 정보를 통해 화자의 발화 유형을 추출하고, 추출된 발화 유형과 화자의 음색을 조합하여, 화자 분석 정보를 생성하는 과정을 설명하는 도면이다.

도 6은 본 발명의 일 실시 예에 따라 도 3의 과정에서 생성된 화자 분석 정보에 기초하여, 제2 언어의 음성을 합성하는 과정을 설명하는 흐름도이다.

도 7 내지 도 9는 본 발명의 일 실시 예에 따라 화자 분석 정보를 이용하여, 제2 언어로된 합성 음성을 생성하는 과정을 설명하는 도면이다.

도 10은 본 발명의 실시 예에 따라 화자의 음성으로부터 화자의 음색 및 음성 스타일(또는 발화 유형)을 반영하여, 제1 언어의 음성을 제2 언어의 음성으로 시청자에게 제공하는 시나리오를 개략적으로 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0015] 이하, 첨부된 도면을 참조하여 본 명세서에 개시된 실시 예를 상세히 설명하되, 도면 부호에 관계없이 동일하거나 유사한 구성요소는 동일한 참조 번호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 이하의 설명에서 사용되는 구성요소에 대한 접미사 "모듈" 및 "부"는 명세서 작성의 용이함만이 고려되어 부여되거나 혼용되는 것으로서, 그 자체로 서로 구별되는 의미 또는 역할을 갖는 것은 아니다. 또한, 본 명세서에 개시된 실시 예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 명세서에 개시된 실시 예의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다. 또한, 첨부된 도면은 본 명세서에 개시된 실시 예를 쉽게 이해할 수 있도록 하기 위한 것일 뿐, 첨부된 도면에 의해 본 명세서에 개시된 기술적 사상이 제한되지 않으며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0016] 제1, 제2 등과 같이 서수를 포함하는 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되지는 않는다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0017] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0018] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함한다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0019] 본 명세서에서 설명되는 단말기에는 휴대폰, 스마트 폰(smart phone), 노트북 컴퓨터(laptop computer), 디지털 방송용 단말기, PDA(personal digital assistants), PMP(portable multimedia player), 네비게이션, 슬레이트 PC(slate PC), 태블릿 PC(tablet PC), 울트라북(ultrabook), 웨어러블 디바이스(wearable device, 예를 들어, 워치형 단말기 (smartwatch), 글래스형 단말기 (smart glass), HMD(head mounted display)) 등과 같은 이동 단말기가 포함될 수 있다.

[0020] 그러나, 본 명세서에 기재된 실시 예에 따른 구성은 이동 단말기에만 적용 가능한 경우를 제외하면, 디지털 TV, 데스크탑 컴퓨터, 디지털사이니지 등과 같은 고정 단말기에도 적용될 수도 있음을 본 기술분야의 당업자라면 쉽게 알 수 있을 것이다.

[0021] 도 1은 본 발명의 일 실시 예에 따른 음성 합성 시스템의 구성을 설명하기 위한 도면이다.

[0022] 본 발명의 실시 예에 따른 음성 합성 시스템(1)은 단말기(10), 음성 합성 장치(20), 음성 데이터 베이스(30) 및 텍스트 데이터 베이스(40)를 포함할 수 있다.

[0023] 음성 합성 장치(20)는 음성 데이터 베이스(30)에 저장된 제1 언어의 음성을 제2 언어의 음성으로 변환할 수 있다.

[0024] 단말기(10)는 음성 합성 장치(20)로부터 제2 언어의 음성으로 변환된 합성 음성을 수신할 수 있다.

[0025] 음성 데이터 베이스(30)는 컨텐트에 포함된 제1 언어의 음성을 저장할 수 있다.

[0026] 음성 데이터 베이스(30)는 제1 언어의 음성이 제2 언어의 음성으로 변환된 합성 음성을 저장할 수 있다.

[0027] 텍스트 데이터 베이스(40)는 제1 언어의 음성에 대응하는 제1 텍스트, 제1 텍스트가 제2 언어로 번역된 제2 텍스트를 저장할 수 있다.

[0028] 음성 데이터 베이스(30) 및 텍스트 데이터 베이스(40)는 음성 합성 장치(20)에 포함될 수도 있다.

[0029] 도 2는 본 발명의 일 실시 예에 따른 음성 합성 장치의 구성을 설명하기 위한 블록도이다.

[0030] 도 2를 참고하면, 음성 합성 장치(20)는 마이크로폰(210), 전처리부(220), 프로세서(230), 스피커(250), 통신부(270) 및 데이터 베이스(290)를 포함할 수 있다.

[0031] 마이크로폰(210)은 화자의 음성을 수신할 수 있다.

[0032] 마이크로폰(510)은 외부의 음향 신호를 전기적인 음성 데이터로 처리할 수 있다.

[0033] 처리된 음성 데이터는 음성 합성 장치(20)에서 수행 중인 기능(또는 실행 중인 응용 프로그램)에 따라 다양하게 활용될 수 있다. 한편, 마이크로폰(210)에는 외부의 음향 신호를 입력 받는 과정에서 발생되는 잡음(noise)을 제거하기 위한 다양한 잡음 제거 알고리즘이 구현될 수 있다.

[0034] 전처리부(220)는 마이크로폰(210)을 통해 수신된 음성 또는 데이터 베이스(290)에 저장된 음성을 전처리 할 수 있다.

[0035] 전처리부(220)는 웨이브 처리부(221), 주파수 처리부(223), 파워 스펙트럼 처리부(225), 음성 텍스트(Speech To Text, STT) 변환부(227)를 포함할 수 있다.

[0036] 웨이브 처리부(221)는 음성의 파형을 추출할 수 있다.

[0037] 주파수 처리부(223)는 음성의 주파수 대역을 추출할 수 있다.

[0038] 파워 스펙트럼 처리부(225)는 음성의 파워 스펙트럼을 추출할 수 있다.

[0039] 파워 스펙트럼은 시간적으로 변동하는 파형이 주어졌을 때, 그 파형에 어떠한 주파수 성분이 어떠한 크기로 포함되고 있는지를 나타내는 파라미터일 수 있다.

[0040] 음성 텍스트(Speech To Text, STT) 변환부(227)는 음성을 텍스트로 변환할 수 있다.

[0041] 음성 텍스트 변환부(227)는 특정 언어의 음성을 해당 언어의 텍스트로 변환할 수 있다.

[0042] 프로세서(230)는 음성 합성 장치(20)의 전반적인 동작을 제어할 수 있다.

[0043] 프로세서(230)는 음성 분석부(231), 텍스트 분석부(232), 특징 클러스터링부(233), 텍스트 매핑부(234) 및 음성 합성부(235)를 포함할 수 있다.

[0044] 음성 분석부(231)는 전처리부(220)에서 전처리된, 음성의 파형, 음성의 주파수 대역 및 음성의 파워 스펙트럼 중 하나 이상을 이용하여, 음성의 특성 정보를 추출할 수 있다.

[0045] 음성의 특성 정보는 화자의 성별 정보, 화자의 목소리(또는 음색, tone), 음의 높낮이, 화자의 말투, 화자의 발화 속도, 화자의 감정 중 하나 이상을 포함할 수 있다.

[0046] 또한, 음성의 특성 정보는 화자의 음색을 더 포함할 수도 있다.

[0047] 텍스트 분석부(232)는 음성 텍스트 변환부(227)에서 변환된 텍스트로부터, 주요 표현 어구를 추출할 수 있다.

[0048] 텍스트 분석부(232)는 변환된 텍스트로부터 어구와 어구 간의 어조가 달라짐을 감지한 경우, 어조가 달라지는 어구를 주요 표현 어구로 추출할 수 있다.

[0049] 텍스트 분석부(232)는 어구와 어구 간의 주파수 대역이 기 설정된 대역 이상 변경된 경우, 어조가 달라진 것으로 판단할 수 있다.

[0050] 텍스트 분석부(232)는 변환된 텍스트의 어구 내에, 주요 단어를 추출할 수도 있다. 주요 단어란 어구 내에 존재하는 명사일 수 있으나, 이는 예시에 불과하다.

[0051] 특징 클러스터링부(233)는 음성 분석부(231)에서 추출된 음성의 특성 정보를 이용하여, 화자의 발화 유형을 분류할 수 있다.

[0052] 특징 클러스터링부(233)는 음성의 특성 정보를 구성하는 유형 항목들 각각에, 가중치를 두어, 화자의 발화 유형을 분류할 수 있다.

[0053] 특징 클러스터링부(233)는 딥러닝 모델의 어텐션(attention) 기법을 이용하여, 화자의 발화 유형을 분류할 수 있다.

[0054] 텍스트 매핑부(234)는 제1 언어로 변환된 텍스트를 제2 언어의 텍스트로 번역할 수 있다.

[0055] 텍스트 매핑부(234)는 제2 언어로 번역된 텍스트를 제1 언어의 텍스트와 매핑 시킬 수 있다.

[0056] 텍스트 매핑부(234)는 제1 언어의 텍스트를 구성하는 주요 표현 어구를 이에 대응하는 제2 언어의 어구에 매핑시킬 수 있다.

[0057] 텍스트 매핑부(234)는 제1 언어의 텍스트를 구성하는 주요 표현 어구에 대응하는 발화 유형을 제2 언어의 어구에 매핑시킬 수 있다. 이는, 제2 언어의 어구에 분류된 발화 유형을 적용시키기 위함이다.

[0058] 음성 합성부(235)는 텍스트 매핑부(234)에서 제2 언어로 번역된 텍스트의 주요 표현 어구에, 특징 클러스터링부(233)에서 분류된 발화 유형 및 화자의 음색을 적용하여, 합성된 음성을 생성할 수 있다.

[0059] 스피커(250)는 합성된 음성을 출력할 수 있다.

[0060] 통신부(270)는 단말기(10) 또는 외부 서버와 유선 또는 무선으로 통신을 수행할 수 있다.

[0061] 데이터 베이스(290)는 도 1에서 설명된 음성 데이터 베이스(30) 및 텍스트 데이터 베이스(40)를 포함할 수 있다.

[0062] 데이터 베이스(290)는 컨텐트에 포함된 제1 언어의 음성을 저장할 수 있다.

[0063] 데이터 베이스(290)는 제1 언어의 음성이 제2 언어의 음성으로 변환된 합성 음성을 저장할 수 있다.

[0064] 데이터 베이스(290)는 제1 언어의 음성에 대응하는 제1 텍스트, 제1 텍스트가 제2 언어로 번역된 제2 텍스트를 저장할 수 있다.

[0065] 한편, 도 2에 포함된 구성 요소들은 도 1에 도시된 단말기(10)에도 포함될 수 있다. 즉, 도 1의 단말기(10)는 음성 합성 장치(20)가 수행하는 동작들을 동일하게 수행할 수 있다.

[0066] 도 3은 본 발명의 일 실시 예에 따른 음성 합성 장치의 동작 방법을 설명하기 위한 흐름도이다.

[0067] 도 3을 참조하면, **음성 합성 장치(20)의 마이크로폰(210)은 화자가 발화한 제1 언어의 음성을 획득한다(S301).**

[0068] 일 실시 예에서, 화자는 영화의 영상에서 등장하는 배우 일 수 있다.

[0069] 마이크로폰(210)은 전체 영상의 한 씬에서 등장하는 배우가 발화한 제1 언어의 음성을 입력 받을 수 있다. 한 씬은 영상을 구성하는 일정 시간 구간을 갖는 영상일 수 있다.

[0070] 여기서, 제1 언어는 영어일 수 있으나, 이는 예시에 불과하다.

[0071] **음성 합성 장치(20)의 프로세서(230)는 획득된 제1 언어의 음성으로부터 화자의 음색(tone) 및 음성의 특성 정보를 추출한다(S303).**

[0072] 일 실시 예에서, 음성 합성 장치(20)의 전처리부(220)는 제1 언어의 음성을 전처리 할 수 있다.

[0073] 전처리부(220)는 제1 언어의 음성으로부터 음성의 파형, 음성의 주파수, 음성의 파워 스펙트럼을 추출할 수 있다.

[0074] 프로세서(230)는 전처리부(220)에 추출된 정보로부터, 화자의 음색을 추출할 수 있다. 예를 들어, 프로세서(230)는 음성의 주파수를 기초로, 화자의 음색을 추출할 수 있다. 구체적으로, 프로세서(230)는 음성의 주파수 대역을 추출하고, 추출된 주파수 대역으로부터 주요 음역대(또는 formant)를 화자의 음색으로 결정할 수 있다.

[0075] 프로세서(230)는 추출된 음성의 파형, 음성의 주파수, 음성의 파워 스펙트럼을 이용하여, 음성의 특성 정보를 추출할 수 있다.

[0076] 음성의 특성 정보는 음성을 발화한 화자의 성별, 화자의 목소리, 음성의 높낮이, 화자의 말투, 음성의 발화 속도, 화자의 감정 중 하나 이상을 포함할 수 있다.

[0077] 후술하겠지만, 화자의 성별, 화자의 목소리, 음성의 높낮이, 화자의 말투, 음성의 발화 속도 및 화자의 감정 각각은 발화의 유형을 나타내는 유형 항목일 수 있다.

[0078] 또한, 본 발명의 실시 예에서, 화자의 음색은 음성의 특성 정보에 포함될 수 있다.

[0079] **음성 합성 장치(20)의 프로세서(230)는 추출된 음성의 특성 정보에 기초하여, 화자의 발화 유형을 분류한다(S305).**

[0080] 일 실시 예에서, 프로세서(230)는 추출된 음성의 특성 정보를 이용하여, 화자의 발화 유형을 분류할 수 있다. 구체적으로, 프로세서(230)는 음성의 특성 정보를 구성하는 유형 항목들 각각에, 가중치를 두어, 화자의 발화 유형을 분류할 수 있다.

[0081] 이에 대해서는 후술한다.

[0082] **음성 합성 장치(20)의 프로세서(230)는 분류된 발화 유형과 화자의 음색을 조합한다(S407).**

[0083] 프로세서(230)는 분류된 발화 유형과 화자의 음색을 조합할 수 있다. 화자의 음색은 추후, 제1 언어의 음성을 제2 언어의 음성으로 합성하는데 사용될 수 있다.

[0084] **음성 합성 장치(20)의 프로세서(230)는 조합 결과를 포함하는 화자 분석 정보를 생성한다(S509).**

[0085] 즉, 화자 분석 정보는 화자의 음색 및 이에 대응하는 발화 유형에 대한 정보를 포함할 수 있다.

[0086] 이하에서는, 도 3의 실시 예를 상세히 설명한다.

[0087] 도 4는 본 발명의 일 실시 예에 따라, 화자의 제1 언어로된 음성으로부터 음성의 특성 정보를 추출하는 예를 설명하는 도면이다.

[0088] 도 4를 참조하면, 마이크로폰(210)은 영화의 한 씬(410)에 등장하는 배우가 발화하는 제1 언어의 음성을 수신할 수 있다.

[0089] 전처리부(220)는 제1 언어의 음성을 전처리하고, 전처리 결과에 기초하여, 제1 언어의 음성에 대응하는 파형(421) 및 파워 스펙트럼(423)을 추출할 수 있다.

[0090] 프로세서(230)는 추출된 파형(421) 및 파워 스펙트럼(423)을 이용하여, 음성의 특성 정보(425)를 추출할 수 있다.

[0091] 음성의 특성 정보(425)는 화자의 성별 정보, 화자의 목소리(또는 음색, tone), 음의 높낮이, 화자의 말투, 화자의 발화 속도, 화자의 감정 중 하나 이상을 포함할 수 있다.

[0092] 일 예로, 프로세서(230)는 일정 시간 동안 발화된 단어의 개수를 이용하여, 화자의 발화 속도를 추출할 수 있다.

[0093] 프로세서(230)는 음성의 피치를 측정하여, 음의 높낮이를 추출할 수 있다.

[0094] 도 5a 및 도 5b는 본 발명의 일 실시 예에 따라, 음성의 특성 정보를 통해 화자의 발화 유형을 추출하고, 추출된 발화 유형과 화자의 음색을 조합하여, 화자 분석 정보를 생성하는 과정을 설명하는 도면이다.

[0095] 먼저, 도 5a를 참조하면, 발화 유형 항목 및 발화 유형 항목에 대응하는 가중치 간의 대응 관계를 나타내는 제1 발화 유형(510)이 도시되어 있다.

[0096] 각 발화 유형은 텍스트를 구성하는 하나의 문장 또는 어구 단위로, 적용될 수 있다.

[0097] 예를 들어, 텍스트에 2개의 문장이 있는 경우, 2개의 문장들 각각에는 서로 다른 발화 유형이 적용될 수 있다.

[0098] 또 다른 예로, 1개의 문장에 2개의 어구가 포함된 경우, 2개의 어구들 각각에는 서로 다른 발화 유형이 적용될 수 있다.

[0099] 도 5a 및 도 5b에서, 가중치는 0 에서 1 사이의 값을 가질 수 있는 값을 가정하여 설명한다.

[0100] 발화 유형 항목이 발화 속도인 경우, 가중치가 1에 가까울수록 발화 속도가 빠르며, 가중치가 0에 가까울수록 발화 속도가 느림을 가정한다.

[0101] 예를 들어, 프로세서(230)는 제1 언어의 음성으로부터 추출된 발화 속도를 통해, 발화 속도 항목의 가중치를

0.2로 설정할 수 있다.

[0102] 발화 유형 항목이 성별인 경우, 가중치가 1에 가까울수록 여성의 음성에 가깝고, 가중치가 0에 가까울수록 남성의 음성에 가까움을 가정한다.

[0103] 프로세서(230)는 제1 언어의 음성으로부터 추출된 성별 정보를 통해, 성별 항목의 가중치를 0.1로 설정할 수 있다.

[0104] 발화 유형 항목이 슬픈 감정인 경우, 가중치가 1에 가까울수록 슬픔의 정도가 크고, 가중치가 0에 가까울수록 슬픔의 정도가 낮음을 가정한다.

[0105] 프로세서(230)는 제1 언어의 음성으로부터 추출된 말투를 통해, 슬픈 감정 항목의 가중치를 0.5로 설정할 수 있다.

[0106] 발화 유형 항목이 음의 높낮이인 경우, 가중치가 1에 가까울수록 음이 높고, 가중치가 0에 가까울수록 음이 낮음을 가정한다.

[0107] 프로세서(230)는 제1 언어의 음성으로부터 추출된 피치를 통해, 음의 높낮이 항목의 가중치를 0.2로 설정할 수 있다.

[0108] 이와 같이, 프로세서(230)는 복수의 발화 유형 항목들 각각에 가중치를 두어, 화자의 제1 발화 유형(510)을 획득할 수 있다.

[0109] 본 발명의 실시 예에 따르면, 복수의 발화 유형 항목들 각각에 가중치를 두어, 화자의 발화 유형을 추출하는 경우, 해당 씬에서 화자의 발화 유형이 보다 정확하게 파악될 수 있다.

[0110] 프로세서(230)는 획득된 제1 발화 유형(510)과 화자의 음색(530)을 조합하여, 화자 분석 정보(550)를 생성할 수 있다.

[0111] 생성된 화자 분석 정보(550)는 추후, 제1 언어의 음성이 제2 언어의 음성으로 변환되는 과정에서 사용될 수 있다.

[0112] 도 5b를 참조하면, 제2 발화 유형(570)에 대한 예가 도시되어 있다.

[0113] 즉, 동일한 화자의 음색이라도, 발화된 문장이나 어구에 따라 서로 다른 발화 유형이 적용될 수 있다.

[0114] 도 5b를 참조하면, 발화 속도 항목에는 0.5의 가중치가, 성별 항목에는 0.1의 가중치가, 슬픈 감정 항목에는 0.3의 가중치가, 음의 높낮이 항목에는 0.7의 가중치가 각각 적용된 제2 발화 유형(570)이 도시되어 있다.

[0115] 이와 같이, 프로세서(230)는 하나의 텍스트 내에 2개의 문장이 있는 경우, 또는 하나의 문장에 2개의 어구가 있는 경우, 각 문장 또는 각 어구를 서로 다른 발화 유형으로 분류할 수 있다.

[0116] 이에 따라, 실제 배우의 목소리 스타일이 번역된 음성으로 그대로 시청자에게 전달될 수 있다.

[0117] 도 6은 본 발명의 일 실시 예에 따라 도 3의 과정에서 생성된 화자 분석 정보에 기초하여, 제2 언어의 음성을 합성하는 과정을 설명하는 흐름도이다.

[0118] 도 6을 참조하면, **음성 합성 장치(20)의 전처리부(220)는 화자의 제1 언어의 음성을 텍스트로 변환한다(S601).**

[0119] 전처리부(220)에 포함된 음성 텍스트 변환부(227)은 제1 언어의 음성을 제1 언어의 텍스트로 변환할 수 있다.

[0120] **음성 합성 장치(20)의 프로세서(230)는 제1 언어로 변환된 텍스트를 제2 언어로 번역한다(S603).**

[0121] 일 실시 예에서, 제1 언어가 영어인 경우, 제2 언어는 한국어 일 수 있다.

[0122] 일 실시 예에서, 프로세서(230)는 제1 언어로 변환된 텍스트와 제2 언어로 번역된 텍스트 간 매핑을 수행할 수 있다. 구체적으로, 프로세서(230)는 제1 언어로 변환된 텍스트로부터 주요 표현들을 추출하고, 추출된 주요 표현들 각각에 대응하는 번역된 표현들 각각을 매핑 시킬 수 있다.

[0123] **음성 합성 장치(20)의 프로세서(230)는 제2 언어로 번역된 텍스트를 화자 분석 정보를 이용하여, 제2 언어의 음성으로 합성한다(S605).**

[0124] 먼저, 프로세서(230)는 도 3의 단계 S309에서 생성된 화자 분석 정보에 포함된 화자의 음색을 이용하여, 제2 언어로 번역된 텍스트에 대응하는 번역 합성 음성을 생성할 수 있다.

[0125] 그 후, 프로세서(230)는 생성된 번역 합성 음성에 화자 분석 정보에 포함된 발화 유형을 반영하여, 최종 합성 음성을 생성할 수 있다.

[0126] 구체적으로, 프로세서(230)는 발화 유형(510)에 포함된 복수의 발화 유형 항목들 각각에 대응하는 가중치를 번역 합성 음성에 반영하여, 최종 합성 음성을 생성할 수 있다.

[0127] **음성 합성 장치(20)의 프로세서(230)는 제2 언어의 음성으로 합성된 합성 음성을 스피커(250)로 출력한다(S607).**

[0128] 일 실시 예에서, 프로세서(230)는 영상의 재생 요청을 수신함에 따라 제2 언어로 합성된 합성 음성을 영상과 함께, 출력할 수 있다.

[0129] 이하에서는 도 6의 실시 예를 상세히 설명한다.

[0130] 도 7 내지 도 9는 본 발명의 일 실시 예에 따라 화자 분석 정보를 이용하여, 제2 언어로된 합성 음성을 생성하는 과정을 설명하는 도면이다.

[0131] 도 7 내지 도 9에서, 제1 언어는 영어이고, 제2 언어는 한국어 임을 가정한다.

[0132] 도 7을 참조하면, 영화의 한 씬(410)이 도시되어 있다.

[0133] 전처리부(220)의 음성 텍스트 변환부(227)는 상기 씬(410)에서 배우가 제1 언어로 발화한 음성을 텍스트(710)로 변환할 수 있다.

[0134] 프로세서(230)는 제1 언어의 음성에 대응하는 텍스트에 복수의 문장들이 포함된 경우, 복수의 문장들 각각으로부터 상기 특성 정보를 추출하고, 추출된 특성 정보에 기초하여, 상기 복수의 문장들 각각에 대응하는 복수의 발화 유형들을 생성할 수 있다.

[0135] 도 7 내지 도 9에서는 주요 표현에 대해 발화 유형을 적용하는 예를 들어 설명한다.

[0136] 실제로, 프로세서(230)는 제2 언어로 번역된 텍스트 전체에 발화 유형을 적용하여, 합성 음성을 생성할 수 있다.

[0137] 제1 언어의 텍스트(710)는 주요 표현들(711, 713)을 포함할 수 있다. 여기서, 주요 표현들(711, 713)은 예시에 불과하다.

[0138] 프로세서(230)에 포함된 텍스트 분석부(232)는 제1 언어의 텍스트(710)로부터 주요 표현들(711, 713)을 추출할 수 있다.

[0139] 텍스트 분석부(232)는 텍스트(710)로부터 어구와 어구 간의 어조가 달라짐을 감지한 경우, 어조가 달라지는 표현을 주요 표현으로 추출할 수 있다.

[0140] 그 후, 프로세서(230)는 도 8에 도시된 바와 같이, 제1 언어의 텍스트(710)를 제2 언어의 텍스트(810)로 번역할 수 있다.

[0141] 프로세서(230)는 제1 언어의 텍스트(710)로부터 추출된 주요 표현들(711, 713) 각각을 제2 언어의 텍스트(810)에 포함된 번역 표현들(811, 813) 각각에 매핑 시킬 수 있다.

[0142] 예를 들어, 프로세서(230)는 제1 주요 표현(711)을 제1 번역 표현(811)에 매핑시킬 수 있고, 제2 주요 표현(713)을 제2 번역 표현(813)에 매핑시킬 수 있다.

[0143] 그 후, 프로세서(230)는 도 9에 도시된 바와 같이, 제2 언어의 텍스트(810)에 포함된 번역 표현들(811, 813) 각각에 화자 분석 정보를 반영할 수 있다.

[0144] 일 실시 예에서, 화자 분석 정보는 하나의 어구 또는 하나의 문장마다 반영될 수 있다.

[0145] 예를 들어, 프로세서(230)는 제1 화자 분석 정보를 제1 번역 표현(811)에 반영하여, 합성 음성을 생성할 수 있고, 제2 화자 분석 정보를 제2 번역 표현(813)에 반영할 수 있다.

[0146] 예를 들어, 제1 화자 분석 정보가 도 5a에 도시된, 제1 발화 유형(510)을 포함하고, 제2 화자 분석 정보가 도 5b에 도시된 제2 발화 유형(530)을 포함하는 경우, 프로세서(230)는 제1 번역 표현(811)에는 제1 발화 유형(510)을 반영하여, 합성 음성을 생성하고, 제2 번역 표현(813)에는 제2 발화 유형(570)을 반영하여, 합성 음성을 생성할 수 있다.

[0147] 프로세서(230)는 제2 언어로 번역된 텍스트(810)에 화자의 음색(530)을 반영하여, 제2 언어로 번역된 번역 합성 음성을 생성할 수 있다.

[0148] 그 후, 프로세서(230)는 생성된 번역 합성 음성에 발화 유형(510)을 반영하여, 최종 합성 음성을 생성할 수 있다. 번역 합성 음성은 동일한 어조로 발화하는 기계음일 수 있다.

[0149] 프로세서(230)는 번역 합성 음성에 발화 유형(510)에 포함된 발화 유형 항목들 각각의 가중치가 반영된 음성 특성을 조합하여, 최종 합성 음성을 생성할 수 있다.

[0150] 이와 같이, 번역 합성 음성에 발화 유형이 반영된 경우, 실제 화자가 번역된 음성으로 발화하는 것과 같은 효과가 주어져, 시청자의 컨텐트 시청 몰입도가 향상될 수 있다.

[0151] 한편, 도 7 내지 도 9에서는 번역된 텍스트 중 주요 표현들에 대해서만, 발화 유형 및 음색이 적용되어, 합성 음성이 생성된 것으로 되어 있으나, 이는 예시에 불과하고, 텍스트 전체에 대해서 발화 유형 및 음색을 적용하여, 합성 음성이 생성될 수 있다.

[0152] 도 10은 본 발명의 실시 예에 따라 화자의 음성으로부터 화자의 음색 및 음성 스타일(또는 발화 유형)을 반영하여, 제1 언어의 음성을 제2 언어의 음성으로 시청자에게 제공하는 시나리오를 개략적으로 설명하는 도면이다.

[0153] 도 10을 참조하면, 각 화자가 제1 언어(1010)로 특정 발화 유형(1030)을 가지고, 발화를 한다.

[0154] 음성 합성 장치(20)는 화자의 음색 및 발화 유형(1030)을 반영한 제2 언어(1050)의 음성을 생성할 수 있다.

[0155] 추후, 시청자는 영화와 같은 컨텐트 재생 시, 성우가 발화한 더빙 음성을 듣지 않고, 실제 배우가 더빙한 것 같은 음성을 들을 수 있다.

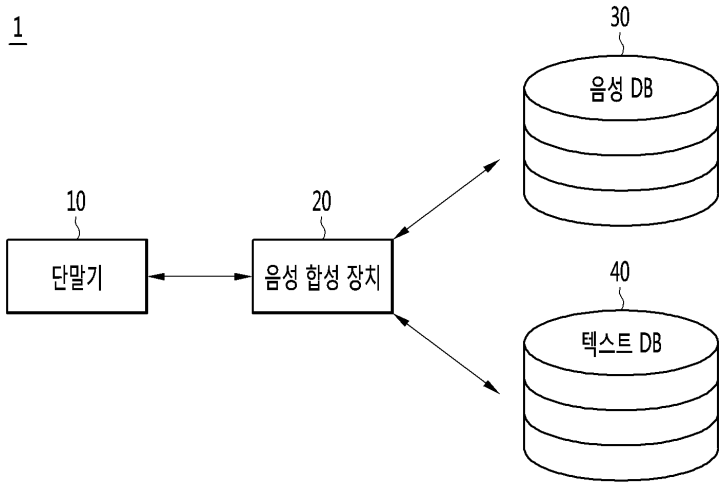
[0156] 이에 따라, 화자의 발화 별 목소리 및 스타일을 반영하는 더빙의 수요가 시청자에게 충족될 수 있다.

[0157] 전술한 본 발명은, 프로그램이 기록된 매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 매체는, 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 읽을 수 있는 매체의 예로는, HDD(Hard Disk Drive), SSD(Solid State Disk), SDD(Silicon Disk Drive), ROM, RAM, CD-ROM, 자기 테이프, 플로피 디스크, 광 데이터 저장 장치 등이 있다. 또한, 상기 컴퓨터는 음성 합성 장치의 프로세서를 포함할 수도 있다.

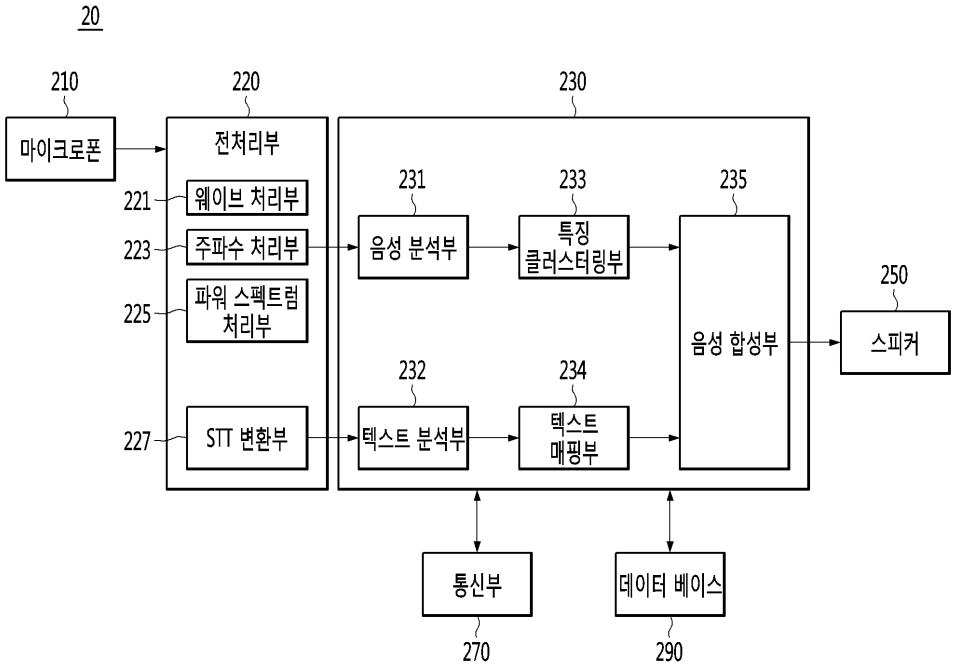
[0158] 따라서, 상기의 상세한 설명은 모든 면에서 제한적으로 해석되어서는 아니되고, 예시적인 것으로 고려되어야 한다. 본 발명의 범위는 첨부된 청구항의 합리적 해석에 의해 결정되어야 하고, 본 발명의 등가적 범위 내에서의 모든 변경은 본 발명의 범위에 포함된다.

도면

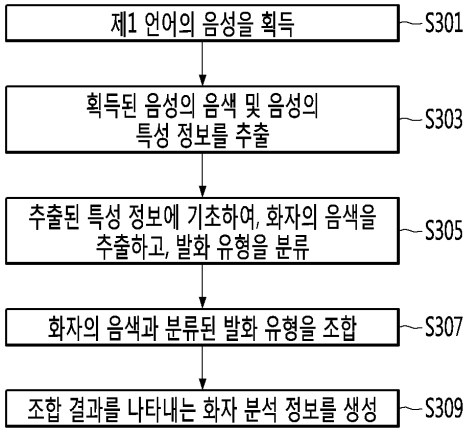
도면1



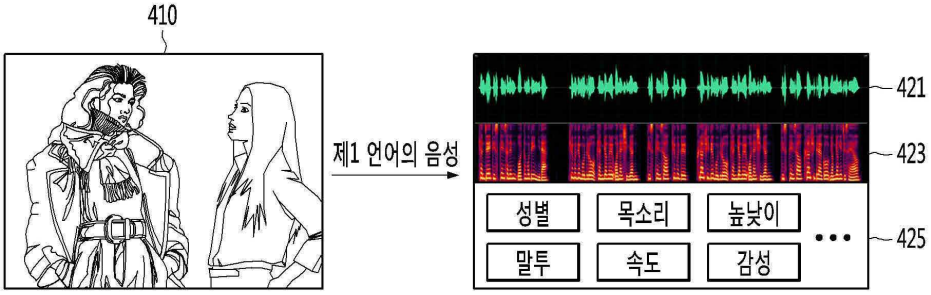
도면2


20
210
마이크로폰
220
전처리부
221
웨이브 처리부
223
주파수 처리부
225
파워 스펙트럼 처리부
227
STT 변환부
230
231
음성 분석부
232
텍스트 분석부
233
특징 클러스터링부
234
텍스트 매핑부
235
음성 합성부
250
스피커
통신부
270
데이터 베이스
290

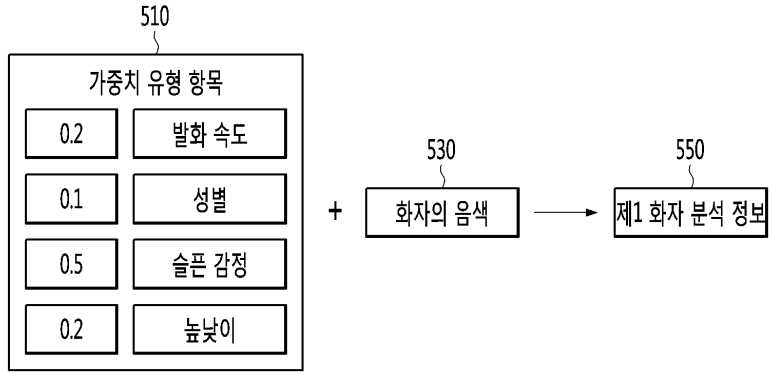
도면3


제1 언어의 음성을 획득
S301
획득된 음성의 음색 및 음성의 특성 정보를 추출
S303
추출된 특성 정보에 기초하여, 화자의 음색을 추출하고, 발화 유형을 분류
S305
화자의 음색과 분류된 발화 유형을 조합
S307
조합 결과를 나타내는 화자 분석 정보를 생성
S309

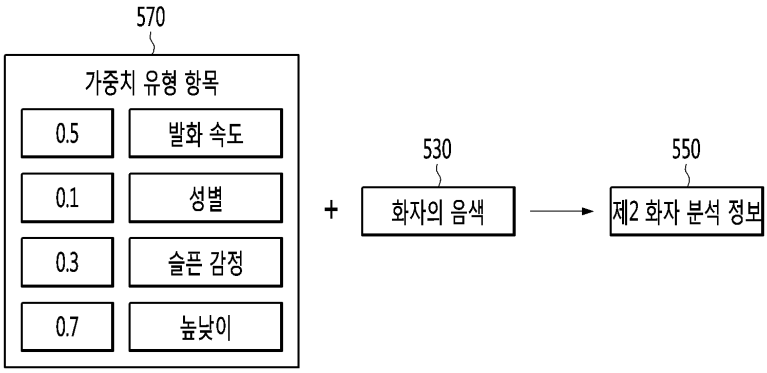
도면4


410
제1 언어의 음성
421
423
성별
목소리
높낮이
말투
속도
감성
425

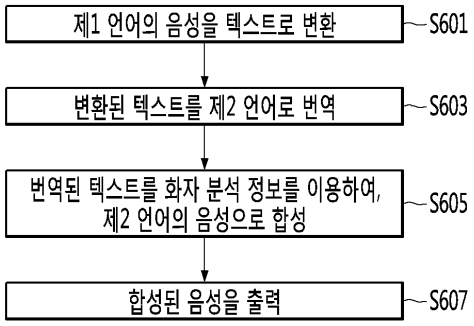
도면5a



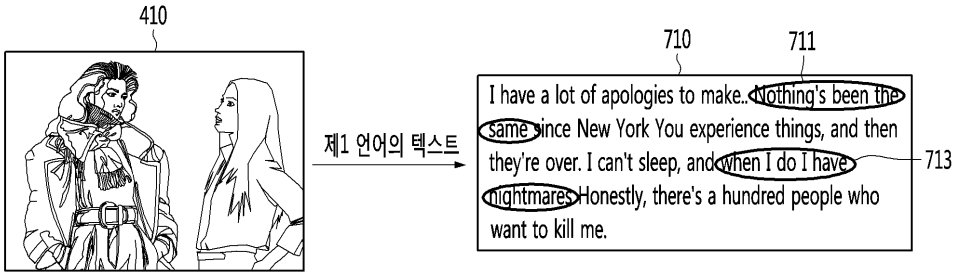
도면5b



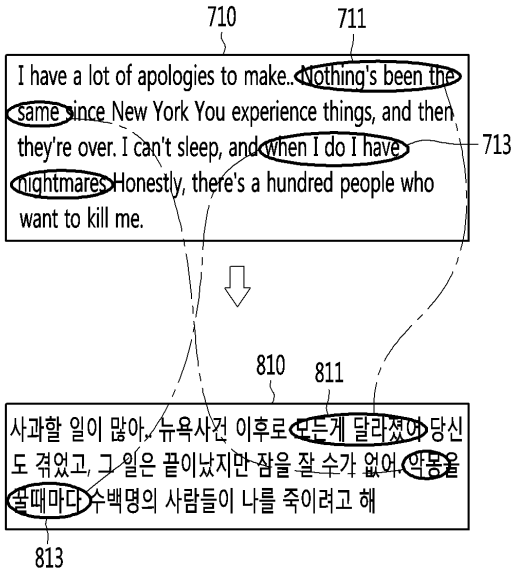
도면6



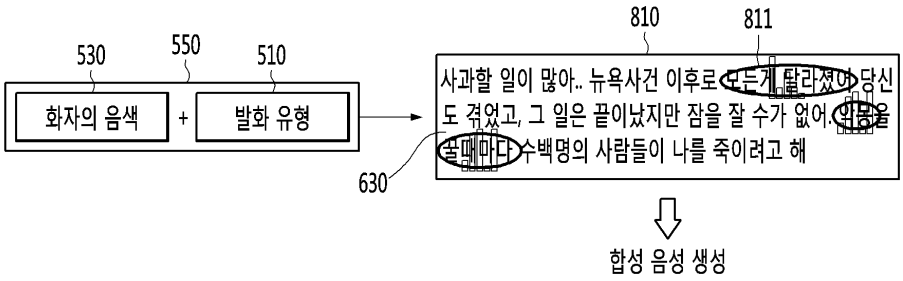
도면7



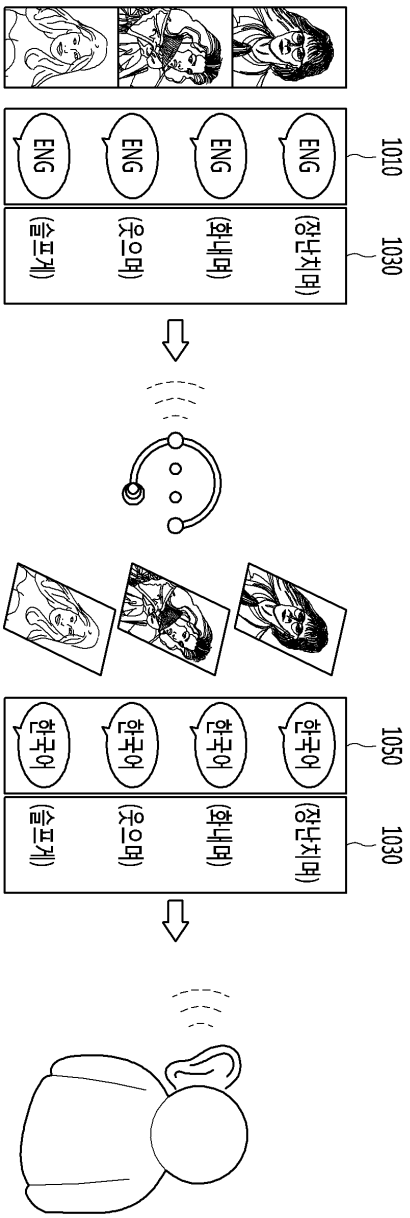
도면8


710
711
I have a lot of apologies to make.. Nothing's been the same since New York You experience things, and then they're over. I can't sleep, and when I do I have nightmares Honestly, there's a hundred people who want to kill me.
713
810
811
사과할 일이 많아.. 뉴욕사건 이후로 모든게 달라졌어 당신도 겪었고, 그 일은 끝이났지만 잠을 잘 수가 없어. 악몽을 꿀때마다 수백명의 사람들이 나를 죽이려고 해
813

도면9


530
550
510
화자의 음색
+
발화 유형
630
810
811
사과할 일이 많아.. 뉴욕사건 이후로 모든게 달라졌어 당신도 겪었고, 그 일은 끝이났지만 잠을 잘 수가 없어. 악몽을 꿀때마다 수백명의 사람들이 나를 죽이려고 해
합성 음성 생성

도면10


ENG
ENG
ENG
ENG
1010
(장난치며)
(화내며)
(웃으며)
(슬프게)
1030
한국어
한국어
한국어
한국어
1050
(장난치며)
(화내며)
(웃으며)
(슬프게)
1030