

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구  
국제사무국

(43) 국제공개일  
2020년 10월 1일 (01.10.2020) WIPO | PCT



(10) 국제공개번호  
WO 2020/196979 A1

(51) 국제특허분류:

G10L 13/033 (2013.01)

G10L 13/08 (2006.01)

G10L 17/02 (2013.01)

G10L 25/51 (2013.01)

G10L 25/30 (2013.01)

(21) 국제출원번호:

PCT/KR2019/004270

(22) 국제출원일:

2019년 4월 10일 (10.04.2019)

(25) 출원언어:

한국어

(26) 공개언어:

한국어

(30) 우선권정보:

10-2019-0033403 2019년 3월 25일 (25.03.2019) KR

(71) 출원인: 한국과학기술원 (KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY) [KR/KR]; 34141 대전시 유성구 대학로 291, Daejeon (KR).

(72) 발명자: 이수영 (LEE, Soo-Young); 34141 대전시 유성구 대학로 291, Daejeon (KR). 김태호 (KIM, Tacho); 34141 대전시 유성구 대학로 291, Daejeon (KR). 신영훈 (SHIN, Young Hoon); 34141 대전시 유성구 대학로 291, Daejeon (KR). 최진국 (CHOI, Shinkook); 34141 대전시 유성구 대학로 291, Daejeon (KR). 김태훈 (KIM, Tae-hun); 34141 대전시 유성구 대학로 291, Daejeon (KR).

라비아잠 (RABIEE, Azam); 34141 대전시 유성구 대학로 291, Daejeon (KR).

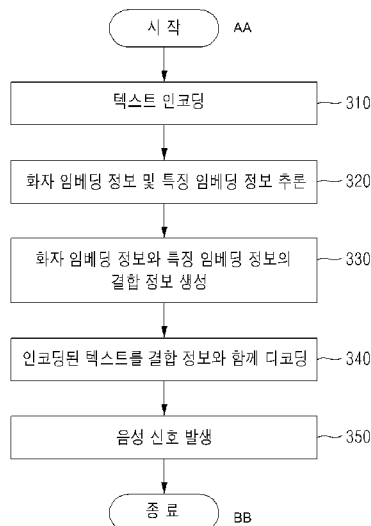
(74) 대리인: 양성보 (YANG, Sungbo); 06099 서울시 강남구 선릉로125길 14 삼성빌딩 2층 피앤티특허법률사무소, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI

(54) Title: ELECTRONIC DEVICE FOR CHARACTERISTICS-CONTROLLABLE SPEECH IMITATION AND OPERATION METHOD THEREOF

(54) 발명의 명칭: 특징 제어 가능 음성 모사를 위한 전자 장치 및 그의 동작 방법



(57) Abstract: An electronic device and an operation method thereof, according to various embodiments, are for characteristics-controllable speech imitation and may comprise: encoding text; inferring embedding information associated with characteristics information and a speech signal of a speaker; decoding the encoded text along with the inferred embedding information; and generating a speech signal.

(57) 요약서: 다양한 실시예들에 따른 전자 장치 및 그의 동작 방법은 특징 제어 가능 음성 모사를 위한 것으로, 텍스트를 인코딩하고, 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하고, 인코딩된 텍스트를 추론된 임베딩 정보와 함께 디코딩하여, 음성 신호를 발생시키도록 구성될 수 있다.

310 ... Encode text  
320 ... Infer speaker embedding information and characteristics embedding information  
330 ... Generate combined information of speaker embedding information and characteristics embedding information  
340 ... Decode encoded text along with combined information  
350 ... Generate speech signal  
AA ... Start  
BB ... End

[다음 쪽 계속]

(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML,  
MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

## 명세서

# 발명의 명칭: 특징 제어 가능 음성 모사를 위한 전자 장치 및 그의 동작 방법

### 기술분야

- [1] 다양한 실시예들은 특징 제어 가능 음성 모사를 위한 전자 장치 및 그의 동작 방법에 관한 것이다.

### 배경기술

- [2] 일반적으로, 음성은 인간의 의사 소통 수단으로서 이용되는 의미 있는 소리이다. 인간과 기계 사이의 음성을 통한 통신 구현에 대한 시도는 과거부터 꾸준히 있어 왔으며, 최근 음성을 효과적으로 처리하는 기술이 실생활에 적용되고 있다. 예를 들면, 음성을 처리하는 기술은 음성 인식(speech recognition), 음성 합성(speech synthesis), 화자 인증(speaker identification and verification) 등을 포함한다. 음성 인식은 화자로부터 발화된 음성을 인식하여 텍스트로 변환하는 기술이고, 음성 합성은 텍스트를 음성으로 변환하는 기술이며, 화자 인증은 음성에 기반하여 화자를 추정하거나 인증하는 기술이다.

### 발명의 상세한 설명

#### 기술적 과제

- [3] 그런데, 상기와 같은 음성 합성에 따르면, 텍스트가 미리 정해진 음성으로만 변환될 뿐이다. 즉 텍스트가 하나의 화자의 음색을 갖는 음성으로만 변환될 뿐이다. 따라서, 텍스트를 다양한 음성으로 변환할 수 있는 방안이 요구된다.

#### 과제 해결 수단

- [4] 다양한 실시예들에 따른 전자 장치의 동작 방법은 특징 제어 가능 음성 모사를 위한 것으로, 텍스트를 인코딩하는 동작, 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하는 동작, 및 상기 인코딩된 텍스트를 상기 추론된 임베딩 정보와 함께 디코딩하여, 음성 신호를 발생시키는 동작을 포함할 수 있다.
- [5] 다양한 실시예들에 따른 전자 장치는 특징 제어 가능 음성 모사를 위한 것으로, 텍스트를 인코딩하는 텍스트 인코더, 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하는 결합부, 상기 인코딩된 텍스트를 상기 추론된 임베딩 정보와 함께 디코딩하는 디코더, 및 상기 디코딩된 텍스트에 대응하는 음성 신호를 발생시키는 보코더를 포함할 수 있다.

#### 발명의 효과

- [6] 다양한 실시예들에 따르면, 전자 장치는 텍스트를 다양한 음성으로 변환할 수 있다. 즉 전자 장치는 화자 음성 신호를 선택적으로 이용하여, 다수의 화자들의 음성을 모사할 수 있다. 아울러, 전자 장치는 화자 음성 신호의 특징을 선택적으로 제어하여, 화자들의 음성을 가변적으로 모사할 수 있다. 예를 들면,

전자 장치는 음성을 다양한 감정으로 표현할 수 있다.

### 도면의 간단한 설명

- [7] 도 1은 다양한 실시예들에 따른 전자 장치를 도시하는 도면이다.
- [8] 도 2는 도 1의 프로세서를 도시하는 도면이다.
- [9] 도 3은 다양한 실시예들에 따른 전자 장치의 동작 방법을 도시하는 도면이다.

### 발명의 실시를 위한 형태

- [10] 이하, 본 문서의 다양한 실시예들이 첨부된 도면을 참조하여 설명된다.
- [11]
- [12] 도 1은 다양한 실시예들에 따른 전자 장치(100)를 도시하는 도면이다.
- [13] 도 1을 참조하면, 다양한 실시예들에 따른 전자 장치(100)는, 입력 모듈(110), 출력 모듈(120), 메모리(130) 또는 프로세서(140) 중 적어도 어느 하나를 포함할 수 있다.
- [14] 입력 모듈(110)은 전자 장치(100)의 구성 요소에 사용될 명령 또는 데이터를 전자 장치(100)의 외부로부터 수신할 수 있다. 입력 모듈(110)은, 사용자가 전자 장치(100)에 직접적으로 명령 또는 데이터를 입력하도록 구성되는 입력 장치 또는 외부 전자 장치와 유선 또는 무선으로 통신하여 명령 또는 데이터를 수신하도록 구성되는 통신 장치 중 적어도 어느 하나를 포함할 수 있다. 예를 들면, 입력 장치는 마이크로폰(microphone), 마우스(mouse), 키보드(keyboard) 또는 카메라(camera) 중 적어도 어느 하나를 포함할 수 있다. 예를 들면, 통신 장치는 유선 통신 장치 또는 무선 통신 장치 중 적어도 어느 하나를 포함하며, 무선 통신 장치는 근거리 통신 장치 또는 원거리 통신 장치 중 적어도 어느 하나를 포함할 수 있다.
- [15] 출력 모듈(120)은 전자 장치(100)의 외부로 정보를 제공할 수 있다. 출력 모듈(120)은 정보를 청각적으로 출력하도록 구성되는 오디오 출력 장치, 정보를 시각적으로 출력하도록 구성되는 표시 장치 또는 외부 전자 장치와 유선 또는 무선으로 통신하여 정보를 전송하도록 구성되는 통신 장치 중 적어도 어느 하나를 포함할 수 있다. 예를 들면, 통신 장치는 유선 통신 장치 또는 무선 통신 장치 중 적어도 어느 하나를 포함하며, 무선 통신 장치는 근거리 통신 장치 또는 원거리 통신 장치 중 적어도 어느 하나를 포함할 수 있다.
- [16] 메모리(130)는 전자 장치(100)의 구성 요소에 의해 사용되는 데이터를 저장할 수 있다. 데이터는 프로그램 또는 이와 관련된 명령에 대한 입력 데이터 또는 출력 데이터를 포함할 수 있다. 예를 들면, 메모리(130)는 휘발성 메모리 또는 비휘발성 메모리 중 적어도 어느 하나를 포함할 수 있다.
- [17] 프로세서(140)는 메모리(130)의 프로그램을 실행하여, 전자 장치(100)의 구성 요소를 제어할 수 있고, 데이터 처리 또는 연산을 수행할 수 있다. 프로세서(140)는 텍스트를 음성 신호로 변환할 수 있다. 여기서, 프로세서(140)는 딥러닝을 위한 신경 회로망을 기반으로, 텍스트를 음성 신호로 변환할 수 있다.

이 때 프로세서(140)는 텍스트를 음성 신호로 변환하는 데 있어서, 특정 화자의 음성을 모사할 수 있다. 예를 들면, 프로세서(140)는 텍스트를 화자의 음색을 갖는 음성 신호로 변환할 수 있다. 이를 통해, 프로세서(140)는, 특정 화자에 의해 텍스트가 발화되는 것과 같이, 음성 신호를 생성할 수 있다. 아울러, 프로세서(140)는 텍스트를 음성 신호로 변환하는 데 있어서, 특정 화자의 음성에서 적어도 하나의 특징을 제어할 수 있다. 이를 위해, 프로세서(140)는 텍스트를 음성 신호로 변환하는 데, 미리 저장된 화자 음성 신호와 미리 저장된 특징 정보를 이용할 수 있다. 이 때 프로세서(140)는 화자 음성 신호에서 특징 정보에 대응하는 특징 요소를 제거하고, 제거된 특징 요소를 대신하여 특징 정보를 적용할 수 있다. 이를 통해, 화자 음성 신호와 특징 정보 간 상관 관계가 제거될 수 있다. 예를 들면, 특징 정보는 감정, 성별 또는 연령 중 적어도 어느 하나를 포함할 수 있다.

[18]

[19] 도 2는 도 1의 프로세서(140)를 도시하는 도면이다.

[20] 도 2를 참조하면, 프로세서(140)는 텍스트 인코더(210), 화자 인코더(220), 특징 인코더(230), 결합부(240), 디코더(250) 또는 보코더(260) 중 적어도 어느 하나를 포함할 수 있다.

[21] 텍스트 인코더(210)는 텍스트를 인코딩할 수 있다. 이 때 텍스트는 입력 모듈(110)을 통해 입력될 수 있다. 일 예로, 텍스트는 입력 장치를 통해 사용자로부터 직접적으로 입력될 수 있다. 다른 예로, 텍스트는 통신 장치를 통해 외부 전자 장치로부터 수신될 수 있다. 또는 텍스트는 메모리(130)에 저장되어 있으며, 프로세서(140)에 의해 메모리(130)로부터 인출될 수 있다.

[22] 화자 인코더(220)는 화자 음성 신호를 인코딩할 수 있다. 여기서, 화자 음성 신호는 메모리(130)에 미리 저장되어 있으며, 가변적 길이를 가질 수 있다. 예를 들면, 화자 음성 신호는 발화 내용을 포함하지 않으며, 화자 음성 신호가 발화 내용을 포함하는 경우, 프로세서(140)가 화자 음성 신호로부터 발화 내용을 제거할 수 있다. 이를 통해, 화자 인코더(220)는 화자 음성 신호로부터 화자 임베딩 정보를 생성할 수 있다. 여기서, 화자 임베딩 정보는 고정된 길이를 가질 수 있다. 이 때 화자 인코더(220)는 RNN(recurrent neural network), GRU(gated recurrent unit), LSTM(long short term memory network) 또는 CNN(convolutional neural network) 중 적어도 어느 하나의 기법을 사용하여, 화자 임베딩 정보를 추론할 수 있다. 어떤 실시예에서는, 화자 인코더(220)는 화자 음성 신호의 화자 인식(speaker identification)을 위한 추가 학습을 수행할 수 있다. 화자 인코더(220)는 LDA(linear discriminate analysis)를 적용하여, 화자 음성 신호의 화자를 구분할 수 있다. 예를 들면, 동일한 화자에 의한 복수 개의 화자 음성 신호들로부터 동일한 화자 임베딩 정보가 생성될 수 있다.

[23] 특징 인코더(230)는 특징 정보를 인코딩할 수 있다. 여기서, 특징 정보는 메모리(130)에 미리 저장되어 있으며, 가변적 길이를 가질 수 있다. 예를 들면,

특징 정보는 이산적인 또는 연속적 특징 변수를 가질 수 있다. 이를 통해, 특징 인코더(230)는 특징 정보로부터 특징 임베딩 정보를 생성할 수 있다. 즉 특징 인코더(230)는 특징 정보와 관련된 이산적 또는 연속적 특징 변수를 기반으로, 특징 임베딩 정보를 추론할 수 있다. 여기서, 특징 임베딩 정보는 고정된 길이를 가질 수 있다. 이 때 특징 인코더(230)는 RNN, GRU, LSTM 또는 CNN 중 적어도 어느 하나의 기법을 사용하여, 특징 임베딩 정보를 추론할 수 있다. 예를 들면, 특징 인코더(230)는 특징 정보와 관련되는 하나의 제어 값의 곱을 사용하여, 특징 정보에 대한 세기 정도를 부여할 수 있으며, 특징 정보와 관련되는 복수 개의 제어 값들들에 대해 복합적인 결과를 획득할 수도 있다.

[24] 결합부(240)는 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론할 수 있다. 이 때 결합부(240)는 화자 임베딩 정보와 특징 임베딩 정보를 결합하여, 임베딩 정보를 생성할 수 있다. 결합부(240)는 가중치 합, 곱셈 또는 신경망(neural network) 중 적어도 어느 하나를 사용하여, 화자 임베딩 정보와 특징 임베딩 정보를 결합할 수 있다. 예를 들면, 신경망은 RNN, GRU, LSTM 또는 CNN 중 적어도 어느 하나를 포함할 수 있다.

[25] 디코더(250)는 인코딩된 텍스트를 임베딩 정보와 함께 디코딩할 수 있다. 즉 디코더(250)가 인코딩된 텍스트와 임베딩 정보를 합성할 수 있다.

[26] 보코더(260)는 디코딩된 텍스트에 대응하는 음성 신호를 발생시킬 수 있다.

[27]

[28] 다양한 실시예들에 따른 전자 장치(100)는 특징 제어 가능 음성 모사를 위한 것으로, 텍스트를 인코딩하는 텍스트 인코더(210), 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하는 결합부(240), 인코딩된 텍스트를 추론된 임베딩 정보와 함께 디코딩하는 디코더(250), 및 디코딩된 텍스트에 대응하는 음성 신호를 발생시키는 보코더(260)를 포함할 수 있다.

[29] 다양한 실시예들에 따르면, 전자 장치(100)는, 화자 음성 신호를 인코딩하여, 화자 임베딩 정보를 추론하는 화자 인코더(220), 및 특징 정보를 인코딩하여, 특징 임베딩 정보를 추론하는 특징 인코더(230)를 더 포함할 수 있다.

[30] 다양한 실시예들에 따르면, 결합부(240)는, 화자 임베딩 정보와 특징 임베딩 정보를 결합하여, 임베딩 정보를 생성할 수 있다.

[31] 다양한 실시예들에 따르면, 결합부(240)는, 가중치 합, 곱셈 또는 신경망 중 적어도 어느 하나를 사용하여, 화자 임베딩 정보와 특징 임베딩 정보를 결합할 수 있다.

[32] 다양한 실시예들에 따르면, 전자 장치(100)는 화자 임베딩 정보와 특징 임베딩 정보 간 상과 관계가 제거되도록, 화자 음성 신호로부터 특징 정보에 대응하는 특징 요소를 제거하는 프로세서(140)를 더 포함할 수 있다.

[33] 다양한 실시예들에 따르면, 특징 정보는 감정, 성별 또는 연령 중 적어도 어느 하나를 포함할 수 있다.

[34] 다양한 실시예들에 따르면, 특징 인코더(230)는 특징 정보와 관련된 이산적

- 또는 연속적 특징 변수를 기반으로, 특징 임베딩 정보를 추론할 수 있다.
- [35] 다양한 실시예들에 따르면, 화자 음성 신호는 가변적 길이를 갖고, 화자 임베딩 정보는 고정된 길이를 가질 수 있다.
- [36] 다양한 실시예들에 따르면, 특징 정보는 가변적 길이를 갖고, 특징 임베딩 정보는 고정된 길이를 가질 수 있다.
- [37]
- [38] 도 3은 다양한 실시예들에 따른 전자 장치(100)의 동작 방법을 도시하는 도면이다.
- [39] 도 3을 참조하면, 전자 장치(100)는 310 동작에서 텍스트를 인코딩할 수 있다. 이 때 프로세서(140)가 텍스트 인코더(210)를 통하여 텍스트를 인코딩할 수 있다. 이 때 텍스트는 입력 모듈(110)을 통해 입력될 수 있다. 일 예로, 텍스트는 입력 장치를 통해 사용자로부터 직접적으로 입력될 수 있다. 다른 예로, 텍스트는 통신 장치를 통해 외부 전자 장치로부터 수신될 수 있다. 또는 텍스트는 메모리(130)에 저장되어 있으며, 프로세서(140)가 메모리(130)로부터 텍스트를 읽어 올 수 있다.
- [40] 전자 장치(100)는 320 동작에서 화자 임베딩 정보와 특징 임베딩 정보를 각각 추론할 수 있다. 메모리(130)에, 적어도 하나의 화자 음성 신호와 특징 정보와 관련된 적어도 하나의 제어 값이 저장되어 있을 수 있다. 여기서, 화자 음성 신호와 특징 정보와 관련된 제어 값은 가변적인 길이를 가질 수 있다. 메모리(130)의 화자 음성 신호는, 화자 음성 신호로부터 생성되는 화자 임베딩 정보의 분산이 최소가 되어야 하는 조건, 화자 음성 신호가 블록 단위로 분할되어 셔플(shuffle)을 통해 생성되는 화자 임베딩 정보가 동일한 값을 가져야 하는 조건 및 동일한 화자에 의한 복수 개의 화자 음성 신호들로부터 동일한 화자 임베딩 정보가 생성되어야 하는 조건을 충족할 수 있다. 프로세서(140)는 화자 음성 신호 중 어느 하나와 특징 정보와 관련된 제어 값 중 적어도 어느 하나를 선택할 수 있다. 프로세서(140)는 화자 인코더(220)를 통해 선택된 화자 음성 신호를 인코딩하여, 화자 임베딩 정보를 추론할 수 있다. 여기서, 화자 임베딩 정보는 고정된 길이를 가질 수 있다. 이 때 프로세서(140)는 화자 음성 신호에서 특징 정보에 대응하는 특징 요소를 제거한 다음, 화자 음성 신호를 인코딩할 수 있다. 이를 통해, 화자 음성 신호와 특징 정보 간 상관 관계가 제거될 수 있다. 화자 인코더(220)는 RNN, GRU, LSTM 또는 CNN 중 적어도 어느 하나의 기법을 사용하여, 화자 임베딩 정보를 추론할 수 있다. 프로세서(140)는 특징 인코더(230)를 통해 선택된 제어 값의 특징 정보를 인코딩하여, 특징 임베딩 정보를 추론할 수 있다. 여기서, 특징 임베딩 정보는 고정된 길이를 가질 수 있다. 특징 인코더(230)는 RNN, GRU, LSTM 또는 CNN 중 적어도 어느 하나의 기법을 사용하여, 특징 임베딩 정보를 추론할 수 있다.
- [41] 전자 장치(100)는 330 동작에서 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론할 수 있다. 이 때 화자 음성 신호에서 제거된 특징 요소를 대신하여

특징 정보가 적용될 수 있다. 즉 전자 장치(100)는 화자 임베딩 정보와 특징 임베딩 정보로부터 임베딩 정보를 생성할 수 있다. 이를 위해, 프로세서(140)가 결합부(240)를 통해 화자 임베딩 정보와 특징 임베딩 정보를 결합하여, 임베딩 정보를 생성할 수 있다. 결합부(240)는 가중치 합, 곱셈 또는 신경망 중 적어도 어느 하나를 사용하여, 화자 임베딩 정보와 특징 임베딩 정보를 결합할 수 있다. 예를 들면, 신경망은 RNN, GRU, LSTM 또는 CNN 중 적어도 어느 하나를 포함할 수 있다.

- [42] 전자 장치(100)는 340 동작에서 인코딩된 텍스트를 임베딩 정보와 함께 디코딩할 수 있다. 프로세서(140)는 디코더(250)를 통해 인코딩된 텍스트를 임베딩 정보와 함께 디코딩할 수 있다.
- [43] 전자 장치(100)는 350 동작에서 디코딩된 텍스트에 대응하는 음성 신호를 발생시킬 수 있다. 프로세서(140)는 보코더(260)를 통해 음성 신호를 발생시킬 수 있다.
- [44]
- [45] 다양한 실시예들에 따른 전자 장치(100)의 동작 방법은 특징 제어 가능 음성 모사를 위한 것으로, 텍스트를 인코딩하는 동작, 화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하는 동작, 및 인코딩된 텍스트를 추론된 임베딩 정보와 함께 디코딩하여, 음성 신호를 발생시키는 동작을 포함할 수 있다.
- [46] 다양한 실시예들에 따르면, 임베딩 정보 추론 동작은, 화자 음성 신호를 인코딩하여, 화자 임베딩 정보를 추론하는 동작, 특징 정보를 인코딩하여, 특징 임베딩 정보를 추론하는 동작, 및 화자 임베딩 정보와 특징 임베딩 정보를 결합하여, 임베딩 정보를 생성하는 동작을 포함할 수 있다.
- [47] 다양한 실시예들에 따르면, 임베딩 정보 생성 동작은, 가중치 합, 곱셈 또는 신경망 중 적어도 어느 하나를 사용하여, 화자 임베딩 정보와 특징 임베딩 정보를 결합하는 동작을 포함할 수 있다.
- [48] 다양한 실시예들에 따르면, 화자 임베딩 정보 추론 동작은, 화자 임베딩 정보와 특징 임베딩 정보 간 상과 관계가 제거되도록, 화자 음성 신호로부터 특징 정보에 대응하는 특징 요소를 제거하는 동작을 포함할 수 있다.
- [49] 다양한 실시예들에 따르면, 특징 정보는 감정, 성별 또는 연령 중 적어도 어느 하나를 포함할 수 있다.
- [50] 다양한 실시예들에 따르면, 특징 임베딩 정보 추론 동작은, 특징 정보와 관련된 이산적 또는 연속적 특징 변수를 기반으로, 특징 임베딩 정보를 추론하는 동작을 포함할 수 있다.
- [51] 다양한 실시예들에 따르면, 화자 음성 신호는 가변적 길이를 갖고, 화자 임베딩 정보는 고정된 길이를 가질 수 있다.
- [52] 다양한 실시예들에 따르면, 특징 정보는 가변적 길이를 갖고, 상기 특징 임베딩 정보는 고정된 길이를 가질 수 있다.
- [53]

- [54] 다양한 실시예들에 따르면, 전자 장치(100)는 텍스트를 다양한 음성으로 변환할 수 있다. 즉 전자 장치(100)는 화자 음성 신호를 선택적으로 이용하여, 다수의 화자들의 음성을 모사할 수 있다. 아울러, 전자 장치(100)는 화자 음성 신호의 특징을 선택적으로 제어하여, 화자들의 음성을 가변적으로 모사할 수 있다. 예를 들면, 전자 장치(100)는 한 화자의 음성을 다양한 감정으로 표현할 수 있다.
- [55]
- [56] 본 문서의 다양한 실시예들 및 이에 사용된 용어들은 본 문서에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 해당 실시 예의 다양한 변경, 균등물, 및/또는 대체물을 포함하는 것으로 이해되어야 한다. 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다. 본 문서에서, "A 또는 B", "A 및/또는 B 중 적어도 하나", "A, B 또는 C" 또는 "A, B 및/또는 C 중 적어도 하나" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다. "제 1", "제 2", "첫째" 또는 "둘째" 등의 표현들은 해당 구성요소들을, 순서 또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다. 어떤(예: 제 1) 구성요소가 다른(예: 제 2) 구성요소에 "(기능적으로 또는 통신적으로) 연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 상기 어떤 구성요소가 상기 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제 3 구성요소)를 통하여 연결될 수 있다.
- [57] 본 문서에서 사용된 용어 "모듈"은 하드웨어, 소프트웨어 또는 펌웨어로 구성된 유닛을 포함하며, 예를 들면, 로직, 논리 블록, 부품, 또는 회로 등의 용어와 상호 호환적으로 사용될 수 있다. 모듈은, 일체로 구성된 부품 또는 하나 또는 그 이상의 기능을 수행하는 최소 단위 또는 그 일부가 될 수 있다. 예를 들면, 모듈은 ASIC(application-specific integrated circuit)으로 구성될 수 있다.
- [58] 본 문서의 다양한 실시예들은 기기(machine)(예: 전자 장치(100))에 의해 읽을 수 있는 저장 매체(storage medium)(예: 메모리(130))에 저장된 하나 이상의 명령어들을 포함하는 소프트웨어로서 구현될 수 있다. 예를 들면, 기기의 프로세서(예: 프로세서(140))는, 저장 매체로부터 저장된 하나 이상의 명령어들 중 적어도 하나의 명령어를 호출하고, 그것을 실행할 수 있다. 이것은 기기가 호출된 적어도 하나의 명령어에 따라 적어도 하나의 기능을 수행하도록 운영되는 것을 가능하게 한다. 하나 이상의 명령어들은 컴파일러에 의해 생성된 코드 또는 인터프리터에 의해 실행될 수 있는 코드를 포함할 수 있다. 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적'은 저장매체가 실재(tangible)하는 장치이고, 신호(signal)(예: 전자기파)를 포함하지 않는다는 것을 의미할 뿐이며, 이 용어는 데이터가 저장매체에 반영구적으로 저장되는 경우와 임시적으로 저장되는

경우를 구분하지 않는다.

- [59] 다양한 실시예들에 따르면, 기술한 구성요소들의 각각의 구성요소(예: 모듈 또는 프로그램)는 단수 또는 복수의 개체를 포함할 수 있다. 다양한 실시예들에 따르면, 전술한 해당 구성요소들 중 하나 이상의 구성요소들 또는 동작들이 생략되거나, 또는 하나 이상의 다른 구성요소들 또는 동작들이 추가될 수 있다. 대체적으로 또는 추가적으로, 복수의 구성요소들(예: 모듈 또는 프로그램)은 하나의 구성요소로 통합될 수 있다. 이런 경우, 통합된 구성요소는 복수의 구성요소들 각각의 구성요소의 하나 이상의 기능들을 통합 이전에 복수의 구성요소들 중 해당 구성요소에 의해 수행되는 것과 동일 또는 유사하게 수행할 수 있다. 다양한 실시예들에 따르면, 모듈, 프로그램 또는 다른 구성요소에 의해 수행되는 동작들은 순차적으로, 병렬적으로, 반복적으로, 또는 휴리스틱하게 실행되거나, 동작들 중 하나 이상이 다른 순서로 실행되거나, 생략되거나, 또는 하나 이상의 다른 동작들이 추가될 수 있다.

## 청구범위

- [청구항 1] 특징 제어 가능 음성 모사를 위한 전자 장치의 동작 방법에 있어서,  
텍스트를 인코딩하는 동작;  
화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하는 동작; 및  
상기 인코딩된 텍스트를 상기 추론된 임베딩 정보와 함께 디코딩하여,  
음성 신호를 발생시키는 동작을 포함하는 방법.
- [청구항 2] 제 1 항에 있어서, 상기 임베딩 정보 추론 동작은,  
상기 화자 음성 신호를 인코딩하여, 화자 임베딩 정보를 추론하는 동작;  
상기 특징 정보를 인코딩하여, 특징 임베딩 정보를 추론하는 동작; 및  
상기 화자 임베딩 정보와 상기 특징 임베딩 정보를 결합하여, 상기 임베딩  
정보를 생성하는 동작을 포함하는 방법.
- [청구항 3] 제 2 항에 있어서, 상기 임베딩 정보 생성 동작은,  
가중치 합, 곱셈 또는 신경망 중 적어도 어느 하나를 사용하여, 상기 화자  
임베딩 정보와 상기 특징 임베딩 정보를 결합하는 동작을 포함하는 방법.
- [청구항 4] 제 2 항에 있어서, 상기 화자 임베딩 정보 추론 동작은,  
상기 화자 임베딩 정보와 상기 특징 임베딩 정보 간 상과 관계가  
제거되도록, 상기 화자 음성 신호로부터 상기 특징 정보에 대응하는 특징  
요소를 제거하는 동작을 포함하는 방법.
- [청구항 5] 제 1 항에 있어서,  
상기 특징 정보는 감정, 성별 또는 연령 중 적어도 어느 하나를 포함하는  
방법.
- [청구항 6] 제 2 항에 있어서, 상기 특징 임베딩 정보 추론 동작은,  
상기 특징 정보와 관련된 이산적 또는 연속적 특징 변수를 기반으로, 상기  
특징 임베딩 정보를 추론하는 동작을 포함하는 방법.
- [청구항 7] 제 2 항에 있어서,  
상기 화자 음성 신호와 상기 특징 정보는 각각 가변적 길이를 갖고,  
상기 화자 임베딩 정보와 상기 특징 임베딩 정보는 각각 고정된 길이를  
갖는 방법.
- [청구항 8] 특징 제어 가능 음성 모사를 위한 전자 장치에 있어서,  
텍스트를 인코딩하는 텍스트 인코더;  
화자 음성 신호 및 특징 정보와 관련된 임베딩 정보를 추론하는 결합부;  
상기 인코딩된 텍스트를 상기 추론된 임베딩 정보와 함께 디코딩하는  
디코더; 및  
상기 디코딩된 텍스트에 대응하는 음성 신호를 발생시키는 보코더를  
포함하는 전자 장치.
- [청구항 9] 제 8 항에 있어서,  
상기 화자 음성 신호를 인코딩하여, 화자 임베딩 정보를 추론하는 화자

인코더; 및

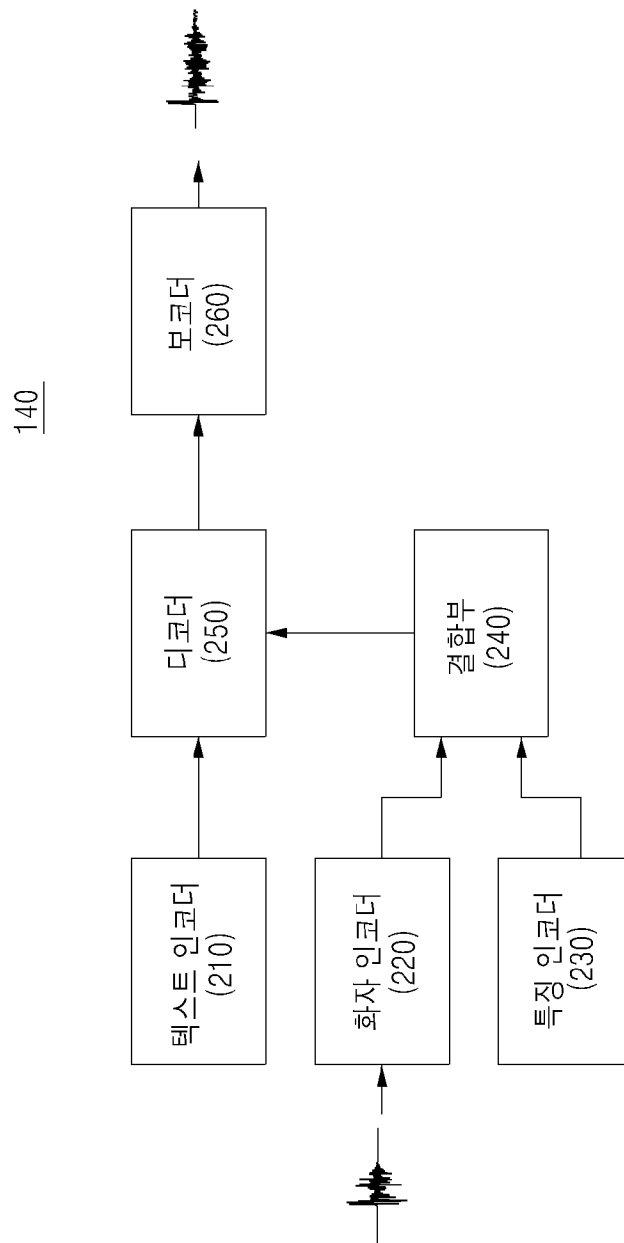
상기 특징 정보를 인코딩하여, 특징 임베딩 정보를 추론하는 특징 인코더를 더 포함하는 전자 장치.

- [청구항 10] 제 9 항에 있어서, 상기 결합부는,  
상기 화자 임베딩 정보와 상기 특징 임베딩 정보를 결합하여, 상기 임베딩 정보를 생성하는 전자 장치.
- [청구항 11] 제 10 항에 있어서, 상기 결합부는,  
가중치 합, 곱셈 또는 신경망 중 적어도 어느 하나를 사용하여, 상기 화자 임베딩 정보와 상기 특징 임베딩 정보를 결합하는 전자 장치.
- [청구항 12] 제 10 항에 있어서,  
상기 화자 임베딩 정보와 상기 특징 임베딩 정보 간 상관 관계가 제거되도록, 상기 화자 음성 신호로부터 상기 특징 정보에 대응하는 특징 요소를 제거하는 프로세서를 더 포함하는 전자 장치.
- [청구항 13] 제 8 항에 있어서,  
상기 특징 정보는 감정, 성별 또는 연령 중 적어도 어느 하나를 포함하는 전자 장치.
- [청구항 14] 제 9 항에 있어서, 상기 특징 인코더는,  
상기 특징 정보와 관련된 이산적 또는 연속적 특징 변수를 기반으로, 상기 특징 임베딩 정보를 추론하는 전자 장치.
- [청구항 15] 제 9 항에 있어서,  
상기 화자 음성 신호와 상기 특징 정보는 각각 가변적 길이를 갖고,  
상기 화자 임베딩 정보와 상기 특징 임베딩 정보는 각각 고정된 길이를 갖는 전자 장치.

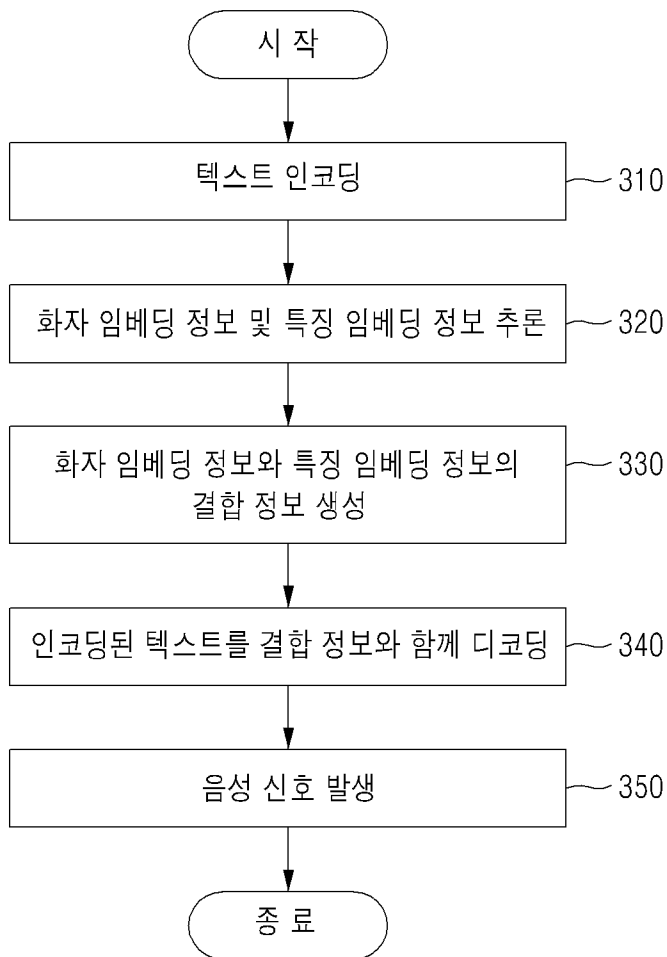
[도1]



[도2]



[도3]



## INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2019/004270

## A. CLASSIFICATION OF SUBJECT MATTER

*G10L 13/033(2013.01)i, G10L 13/08(2006.01)i, G10L 17/02(2013.01)i, G10L 25/51(2013.01)i, G10L 25/30(2013.01)i*

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L 13/033; G06F 3/16; G10L 13/04; G10L 13/08; G10L 19/07; G10L 25/63; G10L 17/02; G10L 25/51; G10L 25/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above

Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) &amp; Keywords: voice, text, speaker, feature, neural network, embedding

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                           | Relevant to claim No. |
|-----------|------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | KR 10-1097186 B1 (MEDIAZEN, INC.) 22 December 2011<br>See paragraphs [0014], [0043] and claim 1.                             | 1,2,4,5,8-10,12,13    |
| Y         |                                                                                                                              | 3,6,7,11,14,15        |
| Y         | KR 10-2019-0016889 A (KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY) 19 February 2019<br>See paragraphs [0066], [0091]. | 3,6,7,11,14,15        |
| A         | KR 10-2015-0017662 A (SAMSUNG ELECTRONICS CO., LTD.) 17 February 2015<br>See claims 1-10 and figures 1-5.                    | 1-15                  |
| A         | JP 2003-233388 A (SHARP CORP.) 22 August 2003<br>See claims 1-4.                                                             | 1-15                  |
| A         | US 2009-0319275 A1 (NODA, Takuya) 24 December 2009<br>See paragraphs [0063]-[0069], claim 1 and figure 4.                    | 1-15                  |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&amp;” document member of the same patent family

Date of the actual completion of the international search

07 JANUARY 2020 (07.01.2020)

Date of mailing of the international search report

07 JANUARY 2020 (07.01.2020)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office  
Government Complex Daejeon Building 4, 189, Cheongsu-ro, Seo-gu,  
Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Telephone No.

**INTERNATIONAL SEARCH REPORT**  
Information on patent family members

International application No.

**PCT/KR2019/004270**

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member | Publication<br>date |
|-------------------------------------------|---------------------|-------------------------|---------------------|
| KR 10-1097186 B1                          | 22/12/2011          | KR 10-2011-0100079 A    | 09/09/2011          |
| KR 10-2019-0016889 A                      | 19/02/2019          | None                    |                     |
| KR 10-2015-0017662 A                      | 17/02/2015          | GB 2516942 A            | 11/02/2015          |
|                                           |                     | GB 2516942 B            | 11/07/2018          |
|                                           |                     | US 2015-0046164 A1      | 12/02/2015          |
| JP 2003-233388 A                          | 22/08/2003          | None                    |                     |
| US 2009-0319275 A1                        | 24/12/2009          | JP 4930584 B2           | 16/05/2012          |
|                                           |                     | US 7987093 B2           | 26/07/2011          |
|                                           |                     | WO 2008-114453 A1       | 25/09/2008          |
|                                           |                     | WO 2008-114453 A9       | 15/10/2009          |

**A. 발명이 속하는 기술분류(국제특허분류(IPC))**

G10L 13/033(2013.01)i, G10L 13/08(2006.01)i, G10L 17/02(2013.01)i, G10L 25/51(2013.01)i, G10L 25/30(2013.01)i

**B. 조사된 분야**

조사된 최소문헌(국제특허분류를 기재)

G10L 13/033; G06F 3/16; G10L 13/04; G10L 13/08; G10L 19/07; G10L 25/63; G10L 17/02; G10L 25/51; G10L 25/30

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 음성(voice), 문자(text), 화자(speaker), 특징(feature), 신경망(neural network), 임베딩(embeding)

**C. 관련 문헌**

| 카테고리* | 인용문헌명 및 관련 구절(해당하는 경우)의 기재                                                        | 관련 청구항                   |
|-------|-----------------------------------------------------------------------------------|--------------------------|
| X     | KR 10-1097186 B1 (미디어젠(주)) 2011.12.22<br>단락 [0014], [0043] 및 청구항 1 참조.            | 1, 2, 4, 5, 8-10, 12, 13 |
| Y     |                                                                                   | 3, 6, 7, 11, 14, 15      |
| Y     | KR 10-2019-0016889 A (한국과학기술원) 2019.02.19<br>단락 [0066], [0091] 참조.                | 3, 6, 7, 11, 14, 15      |
| A     | KR 10-2015-0017662 A (삼성전자주식회사) 2015.02.17<br>청구항 1-10 및 도면 1-5 참조.               | 1-15                     |
| A     | JP 2003-233388 A (SHARP CORP.) 2003.08.22<br>청구항 1-4 참조.                          | 1-15                     |
| A     | US 2009-0319275 A1 (TAKUYA NODA) 2009.12.24<br>단락 [0063]-[0069], 청구항 1 및 도면 4 참조. | 1-15                     |

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

\* 인용된 문헌의 특별 카테고리:

“A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

“D” 본 국제출원에서 출원인이 인용한 문헌

“E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후 “X”에 공개된 선출원 또는 특허 문헌

“L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

“O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

“P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

“T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

“X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

“Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

“&” 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일

2020년 01월 07일 (07.01.2020)

국제조사보고서 발송일

2020년 01월 07일 (07.01.2020)

ISA/KR의 명칭 및 우편주소



대한민국 특허청  
(35208) 대전광역시 서구 청사로 189,  
4동 (둔산동, 정부대전청사)

팩스 번호 +82-42-481-8578

심사관

권성호

전화번호 +82-42-481-3547



| 국제조사보고서에서<br>인용된 특허문헌 | 공개일        | 대응특허문헌               | 공개일        |
|-----------------------|------------|----------------------|------------|
| KR 10-1097186 B1      | 2011/12/22 | KR 10-2011-0100079 A | 2011/09/09 |
| KR 10-2019-0016889 A  | 2019/02/19 | 없음                   |            |
| KR 10-2015-0017662 A  | 2015/02/17 | GB 2516942 A         | 2015/02/11 |
|                       |            | GB 2516942 B         | 2018/07/11 |
|                       |            | US 2015-0046164 A1   | 2015/02/12 |
| JP 2003-233388 A      | 2003/08/22 | 없음                   |            |
| US 2009-0319275 A1    | 2009/12/24 | JP 4930584 B2        | 2012/05/16 |
|                       |            | US 7987093 B2        | 2011/07/26 |
|                       |            | WO 2008-114453 A1    | 2008/09/25 |
|                       |            | WO 2008-114453 A9    | 2009/10/15 |