(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2023 년 4 월 27 일 (27.04.2023)

WIPO | PCT

(10) 국제공개번호
WO 2023/068552 A1

(51) 국제특허분류:
G10L 15/12 (2006.01) *G10L 15/22* (2006.01)
G10L 15/06 (2006.01) *G06F 3/06* (2006.01)
G10L 15/28 (2006.01) *G06F 3/16* (2006.01)

(21) 국제출원번호: PCT/KR2022/013533

(22) 국제출원일: 2022 년 9 월 8 일 (08.09.2022)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2021-0141388 2021 년 10 월 21 일 (21.10.2021) KR
10-2021-0184153 2021 년 12 월 21 일 (21.12.2021) KR

(71) 출원인: 삼성전자주식회사 **(SAMSUNG ELECTRONICS CO., LTD.)** [KR/KR]; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).

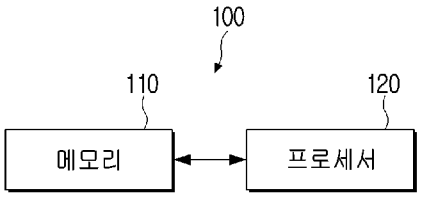
(72) 발명자: **박진환 (PARK, Jinhwan)**; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). **김성수 (KIM, Sungsoo)**; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). **진시첸 (JIN, Sichen)**; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). **박준모 (PARK, Junmo)**; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). **샌디아나다이리아 (SANDHYANA, Dhairya)**; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). **한창우 (HAN, Changwoo)**; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).

(74) 대리인: **김태헌 등 (KIM, Tae-hun** et al.); 06626 서울특별시 서초구 강남대로343 신덕빌딩 9층, 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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: ELECTRONIC DEVICE FOR VOICE RECOGNITION, AND CONTROL METHOD THEREFOR

(54) 발명의 명칭: 음성 인식을 위한 전자 장치 및 그 제어 방법



110 ... Memory
120 ... Processor

(57) Abstract: The present electronic device comprises: a memory for storing a voice recognition model and first recognition information corresponding to a first user voice acquired through the voice recognition model, the voice recognition model including a first network, a second network and a third network; and a processor, which inputs voice data corresponding to a second user voice in the first network so as to acquire a first vector, inputs the first recognition information in the second network for generating a vector on the basis of first weight information, so as to acquire a second vector, inputs the first vector and the second vector in the third network for generating the recognition information on the basis of second weight information, so as to acquire the second recognition information corresponding to the second user voice, wherein at least a part of the second weight information is the same as the first weight information.

WO 2023/068552 A1

[다음 쪽 계속]

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 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 (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

(57) 요약서: 본 전자 장치는 음성 인식 모델 및 음성 인식 모델을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 저장하는 메모리, 상기 음성 인식 모델은 제1 네트워크, 제2 네트워크 및 제3 네트워크를 포함하고, 및 제2 사용자 음성을 에 대응되는 음성 데이터를 상기 제1 네트워크에 입력하여 제1 벡터를 획득하고, 제1 인식 정보를 제1 가중치 정보에 기초하여 벡터를 생성하는 상기 제2 네트워크에 입력하여 제2 벡터를 획득하고, 제1 벡터 및 제2 벡터를 제2 가중치 정보에 기초하여 인식 정보를 생성하는 제3 네트워크에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보를 획득하는 프로세서를 포함하고, 제2 가중치 정보 중 적어도 일부는 제1 가중치 정보와 동일하다.

명세서

발명의 명칭: 음성 인식을 위한 전자 장치 및 그 제어 방법

기술분야

[1] 본 개시는 전자 장치 및 그 제어방법에 관한 것으로, 더욱 상세하게는 음성 인식 모델에 기초하여 사용자 음성에 대응되는 텍스트 정보를 획득하는 전자 장치 및 그 제어방법에 대한 것이다.

배경기술

[2] 음성 인식 모델은 사용자가 발화한 오디오 신호를 텍스트 형태로 출력할 수 있다. 구체적으로, 전자 장치는 오디오 신호를 디지털 신호로 변환하고, 변환된 디지털 신호를 음성 인식 모델에 입력할 수 있다. 여기서, 장치는 음성 인식 모델로부터 사용자의 발화에 대응되는 텍스트 정보를 획득할 수 있다.

[3] 음성 인식 모델을 이용하기 위해서는 2000개 내지 8000개의 단어를 설정하여 사용자 음성을 분석할 수 있다. 또한, 음성 인식 모델은 사용자의 발화를 분석하기 위해 복수의 가중치 또는 파라미터를 이용할 수 있다.

[4] 여기서, 음성 인식 모델에서 이용되는 기 설정된 단어, 가중치 또는 파라미터 등이 저장되는 공간이 확보되지 못하면, 음성 인식 모델이 저장되지 않을 수 있다. 또한, 음성 인식 모델을 이용할 수 있는 메모리가 부족하다면, 처리 속도가 느려지는 문제점이 있다.

[5] 예를 들어, 음성 인식 모델이 사용자의 단말 장치(예를 들어, 스마트폰)에 저장되는 On-device 형태로 구현되는 경우, 메모리 사용량과 저장 공간의 제한이 발생한다는 문제점이 있다.

발명의 상세한 설명

기술적 과제

[6] 본 개시는 상술한 문제를 개선하기 위해 고안된 것으로, 본 개시의 목적은 음성 인식 모델에 이용되는 서로 다른 가중치 정보가 일부 가중치를 공유하는 전자 장치 및 그의 제어 방법을 제공함에 있다.

과제 해결 수단

[7] 본 실시 예에 따라, 전자 장치는 음성 인식 모델 및 상기 음성 인식 모델을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 저장하는 메모리, 상기 음성 인식 모델은 제1 네트워크, 제2 네트워크 및 제3 네트워크를 포함하며, 제2 사용자 음성에 대응되는 음성 데이터를 상기 제1 네트워크에 입력하여 제1 벡터를 획득하고, 상기 제1 인식 정보를 제1 가중치 정보에 기초하여 벡터를 생성하는 상기 제2 네트워크에 입력하여 제2 벡터를 획득하고, 기 제1 벡터 및 상기 제2 벡터를 제2 가중치 정보에 기초하여 인식 정보를 생성하는 상기 제3 네트워크에 입력하여 상기 제2 사용자 음성에 대응되는 제2 인식 정보를 획득하는 프로세서를 포함하고, 상기 제2 가중치 정보 중 적어도 일부는 상기

제1 가중치 정보와 동일하다.

[8] 한편, 상기 음성 인식 모델은 RNN-T(Recurrent Neural Network Transducer) 모델일 수 있다.

[9] 한편, 상기 제1 네트워크는 전사 네트워크(Transcription Network)이고, 상기 제2 네트워크는 예측 네트워크(Prediction Network)이고, 상기 제3 네트워크는 조인트 네트워크(Joint Network)일 수 있다.

[10] 한편, 상기 프로세서는 상기 제2 사용자 음성이 수신되면, 상기 제2 사용자 음성에 대응되는 특징 벡터를 획득하고, 상기 제1 네트워크에 포함된 제1 서브 네트워크는 상기 특징 벡터에 기초하여 상기 제1 벡터를 생성할 수 있다.

[11] 한편, 상기 프로세서는 상기 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득하고, 상기 제2 네트워크에 포함된 제2 서브 네트워크는 상기 원-핫 벡터 및 상기 제1 가중치 정보에 기초하여 상기 제2 벡터를 생성할 수 있다.

[12] 한편, 상기 프로세서는 제3 네트워크에 포함된 제3 서브 네트워크는 상기 제1 벡터 및 상기 제2 벡터에 기초하여 제3 벡터를 획득하고, 상기 제3 네트워크는 상기 제3 벡터 및 상기 제2 가중치 정보에 기초하여 상기 제2 인식 정보를 생성할 수 있다.

[13] 한편, 상기 제1 가중치 정보는 기 설정된 개수의 서브 워드에 대응되는 적어도 하나의 제1 가중치를 포함하고, 상기 제2 가중치 정보는 상기 적어도 하나의 제1 가중치 및 적어도 하나의 추가 가중치를 포함하고, 상기 적어도 하나의 제1 가중치는 상기 메모리의 제1 영역에 저장되고, 상기 적어도 하나의 추가 가중치는 상기 메모리의 제2 영역에 저장되고, 상기 프로세서는 상기 제1 영역에 저장된 상기 적어도 하나의 제1 가중치 및 상기 제2 영역에 저장된 상기 적어도 하나의 가중치를 상기 제2 가중치 정보로써 이용할 수 있다.

[14] 한편, 상기 적어도 하나의 추가 가중치는 상기 제2 사용자 음성에 대응되는 상기 기 설정된 개수의 서브 워드가 존재하지 않는 경우 이용되는 가중치이고, 상기 적어도 하나의 제1 가중치의 차원(dimension)은 상기 적어도 하나의 추가 가중치의 차원(dimension)에 대응될 수 있다.

[15] 한편, 상기 제1 가중치 정보는 상기 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트(gradient), 상기 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트(gradient) 및 학습률(Learning rate)에 기초하여 학습되고, 상기 제2 가중치 정보는 상기 학습된 제1 가중치 정보에 기초하여 결정될 수 있다.

[16] 한편, 상기 제1 가중치 정보 및 상기 제2 가중치 정보 각각은 제1 서브 가중치 정보 및 제2 서브 가중치 정보의 평균값에 기초하여 학습되고, 상기 제1 서브 가중치 정보는 상기 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트 및 학습률에 기초하여 결정되고, 상기 제2 서브 가중치 정보는 상기 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트 및 상기 학습률에 기초하여 결정될 수 있다.

[17] 본 실시 예에 따른, 음성 인식 모델 및 상기 음성 인식 모델을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 저장하는 전자 장치의 제어 방법은 상기 음성 인식 모델은 제1 네트워크, 제2 네트워크 및 제3 네트워크를 포함하고, 제2 사용자 음성에 대응되는 음성 데이터를 상기 제1 네트워크에 입력하여 제1 벡터를 획득하는 단계, 상기 제1 인식 정보를 제1 가중치 정보에 기초하여 벡터를 생성하는 상기 제2 네트워크에 입력하여 제2 벡터를 획득하는 단계 및 상기 제1 벡터 및 상기 제2 벡터를 제2 가중치 정보에 기초하여 인식 정보를 생성하는 상기 제3 네트워크에 입력하여 상기 제2 사용자 음성에 대응되는 제2 인식 정보를 획득하는 단계를 포함하고, 상기 제2 가중치 정보 중 적어도 일부는 상기 제1 가중치 정보와 동일하다.

[18] 한편, 상기 음성 인식 모델은 RNN-T(Recurrent Neural Network Transducer) 모델일 수 있다.

[19] 한편, 상기 제1 네트워크는 전사 네트워크(Transcription Network)이고, 상기 제2 네트워크는 예측 네트워크(Prediction Network)이고, 상기 제3 네트워크는 조인트 네트워크(Joint Network)일 수 있다.

[20] 한편, 상기 제1 벡터를 획득하는 단계는 상기 제2 사용자 음성이 수신되면, 상기 제2 사용자 음성에 대응되는 특징 벡터를 획득하고, 상기 제1 네트워크에 포함된 제1 서브 네트워크는 상기 특징 벡터에 기초하여 상기 제1 벡터를 생성할 수 있다.

[21] 한편, 상기 제2 벡터를 획득하는 단계는 상기 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득하고, 상기 제2 네트워크에 포함된 제2 서브 네트워크는 상기 원-핫 벡터 및 상기 제1 가중치 정보에 기초하여 상기 제2 벡터를 생성할 수 있다.

도면의 간단한 설명

[22] 본 개시의 특정 실시 예의 상술한 내용 및 다른 내용, 특징, 이점은 아래의 도면과 함께 기재된 구체적인 설명으로부터 명백해질 수 있다.

[23] 도 1은 본 개시의 일 실시 예에 따른 전자 장치를 도시한 블록도이다.

[24] 도 2는 복수의 네트워크로 구성된 음성 인식 모델을 설명하기 위한 도면이다.

[25] 도 3은 이전 출력값에 기초하여 인식 정보를 획득하는 음성 인식 모델을 설명하기 위한 도면이다.

[26] 도 4는 복수의 이전 출력값들에 기초하여 인식 정보를 획득하는 음성 인식 모델을 설명하기 위한 도면이다.

[27] 도 5는 제1 가중치 정보 및 제2 가중치 정보가 동일하지 않은 실시 예의 가중치 정보 저장 방식을 설명하기 위한 도면이다.

[28] 도 6는 제1 가중치 정보 및 제2 가중치 정보가 동일하지 않은 실시 예의 가중치 정보 구성을 설명하기 위한 도면이다.

[29] 도 7은 제1 가중치 정보 및 제2 가중치 정보가 일부 동일한 실시 예의 가중치

정보 저장 방식을 설명하기 위한 도면이다.

[30] 도 8은 제1 가중치 정보 및 제2 가중치 정보가 일부 동일한 실시 예의 가중치 정보 구성을 설명하기 위한 도면이다.

[31] 도 9는 음성 인식 모델을 이용하여 사용자 음성에 대응되는 인식 정보를 획득하는 동작을 설명하기 위한 흐름도이다.

[32] 도 10은 제1 사용자 음성 및 제2 사용자 음성에 기초하여 인식 정보를 획득하는 동작을 설명하기 위한 흐름도이다.

[33] 도 11은 제1 벡터를 획득하는 구체적인 동작을 설명하기 위한 흐름도이다.

[34] 도 12는 제2 벡터를 획득하는 구체적인 동작을 설명하기 위한 흐름도이다.

[35] 도 13은 제3 벡터를 획득하는 구체적인 동작을 설명하기 위한 흐름도이다.

[36] 도 14는 일 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 도면이다.

[37] 도 15는 일 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 흐름도이다.

[38] 도 16은 다른 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 도면이다.

[39] 도 17은 다른 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 흐름도이다.

[40] 도 18은 본 개시의 일 실시 예에 따른 전자 장치의 제어 방법을 설명하기 위한 흐름도이다.

발명의 실시를 위한 형태

[41] 이하에서는 첨부 도면을 참조하여 본 개시를 상세히 설명한다.

[42] 본 개시의 실시 예에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 개시의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

[43] 본 명세서에서, "가진다," "가질 수 있다," "포함한다," 또는 "포함할 수 있다" 등의 표현은 해당 특징(예: 수치, 기능, 동작, 또는 부품 등의 구성요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[44] A 또는/및 B 중 적어도 하나라는 표현은 "A" 또는 "B" 또는 "A 및 B" 중 어느 하나를 나타내는 것으로 이해되어야 한다.

[45] 본 명세서에서 사용된 "제1," "제2," "첫째," 또는 "둘째,"등의 표현들은 다양한 구성요소들을, 순서 및/또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.

[46] 어떤 구성요소(예: 제1 구성요소)가 다른 구성요소(예: 제2 구성요소)에 "(기능적으로 또는 통신적으로) 연결되어((operatively or communicatively) coupled with/to)" 있다거나 "접속되어(connected to)" 있다고 언급된 때에는, 어떤 구성요소가 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제3 구성요소)를 통하여 연결될 수 있다고 이해되어야 할 것이다.

[47] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "구성되다" 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.

[48] 본 개시에서 "모듈" 혹은 "부"는 적어도 하나의 기능이나 동작을 수행하며, 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다. 또한, 복수의 "모듈" 혹은 복수의 "부"는 특정한 하드웨어로 구현될 필요가 있는 "모듈" 혹은 "부"를 제외하고는 적어도 하나의 모듈로 일체화되어 적어도 하나의 프로세서(미도시)로 구현될 수 있다.

[49] 본 명세서에서, 사용자라는 용어는 전자 장치를 사용하는 사람 또는 전자 장치를 사용하는 장치(예: 인공지능 전자 장치)를 지칭할 수 있다.

[50] 이하 첨부된 도면들을 참조하여 본 개시의 일 실시 예를 보다 상세하게 설명한다.

[51] 도 1은 본 개시의 일 실시 예에 따른 전자 장치(100)를 도시한 블록도이다.

[52] 도 1을 참조하면, 전자 장치(100)는 메모리(110) 및 프로세서(120)를 포함할 수 있다.

[53] 본 명세서의 다양한 실시 예들에 따른 전자 장치(100)는, 예를 들면, 스마트폰, 태블릿 PC, 이동 전화기, 데스크탑 PC, 랩탑 PC, PDA, PMP(portable multimedia player) 중 적어도 하나를 포함할 수 있다. 어떤 실시 예들에서, 전자 장치(100)는, 예를 들면, 텔레비전, DVD(digital video disk) 플레이어, 미디어 박스(예: 삼성 HomeSyncTM, 애플TVTM, 또는 구글 TVTM)중 적어도 하나를 포함할 수 있다.

[54] 메모리(110)는 프로세서(120)에 포함된 롬(ROM)(예를 들어, EEPROM(electrically erasable programmable read-only memory)), 램(RAM) 등의 내부 메모리로 구현되거나, 프로세서(120)와 별도의 메모리로 구현될 수도 있다. 이 경우, 메모리(110)는 데이터 저장 용도에 따라 전자 장치(100)에 임베디드된 메모리 형태로 구현되거나, 전자 장치(100)에 탈부착이 가능한 메모리 형태로 구현될 수도 있다. 예를 들어, 전자 장치(100)의 구동을 위한 데이터의 경우 전자 장치(100)에 임베디드된 메모리에 저장되고, 전자 장치(100)의 확장 기능을 위한 데이터의 경우 전자 장치(100)에 탈부착이 가능한 메모리에 저장될 수 있다.

[55] 한편, 전자 장치(100)에 임베디드된 메모리의 경우 휘발성 메모리(예: DRAM(dynamic RAM), SRAM(static RAM), 또는 SDRAM(synchronous dynamic

RAM) 등), 비휘발성 메모리(non-volatile Memory)(예: OTPROM(one time programmable ROM), PROM(programmable ROM), EPROM(erasable and programmable ROM), EEPROM(electrically erasable and programmable ROM), mask ROM, flash ROM, 플래시 메모리(예: NAND flash 또는 NOR flash 등), 하드 드라이브, 또는 솔리드 스테이트 드라이브(solid state drive(SSD)) 중 적어도 하나로 구현되고, 전자 장치(100)에 탈부착이 가능한 메모리의 경우 메모리 카드(예를 들어, CF(compact flash), SD(secure digital), Micro-SD(micro secure digital), Mini-SD(mini secure digital), xD(extreme digital), MMC(multi-media card) 등), USB 포트에 연결 가능한 외부 메모리(예를 들어, USB 메모리) 등과 같은 형태로 구현될 수 있다.

[56] 프로세서(120)는 전자 장치(100)의 전반적인 제어 동작을 수행할 수 있다. 구체적으로, 프로세서(120)는 전자 장치(100)의 전반적인 동작을 제어하는 기능을 한다.

[57] 프로세서(120)는 디지털 신호를 처리하는 디지털 시그널 프로세서(digital signal processor(DSP), 마이크로 프로세서(microprocessor), TCON(Time controller)으로 구현될 수 있다. 다만, 이에 한정되는 것은 아니며, 중앙처리장치(central processing unit(CPU)), MCU(Micro Controller Unit), MPU(micro processing unit), 컨트롤러(controller), 어플리케이션 프로세서(application processor(AP)), GPU(graphics-processing unit) 또는 커뮤니케이션 프로세서(communication processor(CP)), ARM 프로세서 중 하나 또는 그 이상을 포함하거나, 해당 용어로 정의될 수 있다. 또한, 프로세서(120)는 프로세싱 알고리즘이 내장된 SoC(System on Chip), LSI(large scale integration)로 구현될 수도 있고, FPGA(Field Programmable gate array) 형태로 구현될 수도 있다. 또한, 프로세서(120)는 메모리(110)에 저장된 컴퓨터 실행가능 명령어(computer executable instructions)를 실행함으로써 다양한 기능을 수행할 수 있다.

[58] 한편, 메모리(110)는 본 개시의 일 실시 예에 따라 도 2에 개시된 음성 인식 모델(200)을 저장할 수 있다. 음성 인식 모델(200)은 신경망 네트워크들과 같은 복수의 네트워크를 포함할 수 있다. 여기서, 음성 인식 모델(200)는 제1 네트워크(210), 제2 네트워크(220) 및 제3 네트워크(230)를 포함할 수 있다. 음성 인식 모델(200)은 오디오 신호 또는 사용자 음성에 대응되는 음성 데이터를 입력 데이터로 수신할 수 있고 사용자 음성에 대응되는 인식 정보(또는 텍스트 정보)를 출력 데이터로서 생성할 수 있다. 인식 정보는 사용자 음성에 대응되는 텍스트 정보를 의미할 수 있다.

[59] 음성 인식 모델(200)에 포함되는 복수의 네트워크와 관련된 구체적인 설명은 도 2에서 기재한다.

[60] 한편, 프로세서(120)는 제1 사용자 음성에 대응되는 적어도 하나의 제1 오디오 신호들을 획득할 수 있다. 그리고, 프로세서(120)는 음성 인식 모델(200)에 기초하여 제1 사용자 음성에 대응되는 제1 인식 정보를 획득할 수

있다.프로세서(120)는 제1 오디오 신호 또는 제1 사용자 음성에 대응되는 다른 제1 음성 데이터를 입력 데이터로써 음성 인식 모델(200)에 입력하고, 음성 인식 모델(200)로부터 제1 사용자 음성에 대응되는 제1 인식 정보를 출력 데이터로서 획득할 수 있다. 프로세서(120)는 음성 인식 모델(200)을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 메모리(110)에 저장할 수 있다. 따라서, 메모리(110)는 제1 사용자 음성에 대응되는 제1 인식 정보를 저장할 수 있다.

[61] 한편, 프로세서(120)는 제2 사용자 음성을 복수의 네트워크 중 제1 네트워크(210)에 입력하여 제1 벡터를 획득하고, 제1 인식 정보를 복수의 네트워크 중 제1가중치 정보를 포함하는 제2 네트워크(220)에 입력하여 제2 벡터를 획득하고, 제1 벡터 및 제2 벡터를 복수의 네트워크 중 제2 가중치 정보를 포함하는 제3 네트워크(230)에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보를 획득할 수 있다. 여기서, 제2 가중치 정보 중 적어도 일부는 제1 가중치 정보와 동일한 정보일 수 있다.

[62] 여기서, 프로세서(120)는 제1 사용자 음성과 상이한 제2 사용자 음성을 획득할 수 있다. 프로세서(120)는 음성 인식 모델(200)을 통해 제2 사용자 음성에 대응되는 제2 인식 정보를 획득할 수 있다. 구체적으로, 프로세서(120)는 제2 사용자 음성을 입력 데이터로서 음성 인식 모델(200)에 입력하고, 음성 인식 모델(200)로부터 제2 사용자 음성에 대응되는 제2 인식 정보를 출력 데이터로서 획득할 수 있다.

[63] 여기서, 프로세서(120)는 제1 네트워크(210)를 통해 제1 벡터를 획득할 수 있다. 구체적으로, 프로세서(120)는 제2 사용자 음성을 제1 네트워크(210)에 입력하여 제1 벡터를 획득할 수 있다. 제1 벡터는 사용자 음성(제2 사용자 음성)에 기초하여 획득되는 히든 벡터를 의미할 수 있다. 제1 벡터와 관련된 구체적인 설명은 도 3의 수학식(211-1)을 통해 기재한다.

[64] 여기서, 프로세서(120)는 제2 네트워크(220)를 통해 제2 벡터를 획득할 수 있다. 구체적으로, 프로세서(120)는 제1 사용자 음성에 대응되는 제1 인식 정보를 제2 네트워크(220)에 입력하여 제2 벡터를 획득할 수 있다. 제2 네트워크(220)는 제1 가중치 정보를 포함(또는 저장)할 수 있다. 프로세서(120)는 제1 인식 정보 및 제1 가중치 정보에 기초하여 제2 벡터를 획득할 수 있다. 제2 벡터는 이전 출력 결과(제1 인식 정보)에 기초하여 획득되는 히든 벡터를 의미할 수 있다. 여기서, 제2 벡터와 관련된 구체적인 설명은 도 3의 수학식(221-1) 및 수학식(222-1)을 통해 기재한다.

[65] 여기서, 프로세서(120)는 제3 네트워크(230)를 통해 제2 사용자 음성에 대응되는 제2 인식 정보를 획득할 수 있다. 구체적으로, 프로세서(120)는 제1 벡터 및 제2 벡터를 제3 네트워크(230)에 입력하여 제2 인식 정보를 획득할 수 있다. 프로세서(120)는 제3 네트워크(230)를 통해 제1 벡터 및 제2 벡터에 기초하여 제3 벡터를 획득할 수 있다. 제3 벡터는 제1 벡터 및 제2 벡터를 결합한 벡터를 의미할 수 있다. 그리고, 제3 네트워크(230)는 제2 가중치 정보를

포함(또는 저장)할 수 있다. 프로세서(120)는 제2 가중치 정보 및 제3 벡터에 기초하여 제2 인식 정보를 획득할 수 있다. 여기서, 제2 인식 정보를 획득하는 동작은 도3의 수학식(231) 및 수학식(232)을 통해 기재한다.

[66] 여기서, 제1 가중치 정보 및 제2 가중치 정보는 동일한 정보를 포함할 수 있다. 제1 가중치 정보에 포함된 정보가 제2 가중치 정보에 포함될 수 있다. 제2 가중치 정보는 제1 가중치 정보에 포함된 정보를 포함할 수 있다. 예를 들어, 제1 가중치 정보에 포함된 가중치가 제2 가중치 정보에 포함될 수 있다. 또한, 제2 가중치 정보는 제1 가중치 정보에 포함된 가중치를 포함할 수 있다.

[67] 다만, 제2 가중치 정보는 제1 가중치 정보 이외에 추가 가중치 정보를 더 포함할 수 있다. 예를 들어, 제2 가중치 정보는 제1 가중치 정보에 포함된 가중치 및 추가 가중치를 포함할 수 있다.

[68] 따라서, 프로세서(120)는 제1 가중치 정보에 포함된 가중치를 메모리(110)의 제1 영역에 저장하고 추가 가중치를 메모리(110)의 제2 영역에 저장할 수 있다. 그리고, 프로세서(120)는 메모리(110)의 제1 영역에 저장된 가중치를 제1 가중치 정보로서 이용할 수 있다. 그리고, 프로세서(120)는 메모리(110)의 제1 영역에 저장된 가중치 및 메모리(110)의 제2 영역에 저장된 추가 가중치를 제2 가중치 정보로서 이용할 수 있다. 제1 가중치 정보 및 제2 가중치 정보에 중복되는 가중치를 하나의 영역에 저장함으로써 저장 공간을 효율적으로 이용할 수 있다. 메모리의 영역과 관련된 설명은 도 5 내지 도 8에서 기재한다.

[69] 여기서, 가중치 정보는 파라미터 정보 또는 임베딩 등으로 기재될 수 있다. 예를 들어, 제1 가중치 정보는 제1 파라미터 정보 또는 제1 임베딩으로 기재될 수 있다.

[70] 한편, 음성 인식 모델(200)은 RNN-T(Recurrent Neural Network Transducer) 모델일 수 있다.

[71] 여기서, RNN-T 모델은 사용자 음성이 계속하여 입력되는 중간 과정에서도 예측 동작을 수행하는 실시간 음성 인식 모델일 수 있다. 구체적으로, RNN-T 모델은 전사 네트워크(Transcription Network), 예측 네트워크(Prediction Network) 및 조인트 네트워크(Joint Network)를 포함할 수 있다.

[72] 여기서, 전사 네트워크는 실시간 사용자 음성에 대응되는 벡터를 획득하는 네트워크일 수 있다. 예측 네트워크는 이전 사용자 음성에 대응되는 벡터를 획득하는 네트워크일 수 있다. 조인트 네트워크는 전사 네트워크에서 출력된 벡터와 예측 네트워크에서 출력된 벡터를 결합하는 네트워크일 수 있다.

[73] 한편, 제1 네트워크(210)는 전사 네트워크(Transcription Network)이고, 제2 네트워크(220)는 예측 네트워크(Prediction Network)이고, 제3 네트워크(230)는 조인트 네트워크(Joint Network)일 수 있다.

[74] 한편, 프로세서(120)는 제2 사용자 음성이 수신되면, 제2 사용자 음성에 대응되는 특징 벡터를 획득하고, 제2 사용자 음성에 대응되는 특징 벡터 및 제1 네트워크(210)에 포함된 제1 서브 네트워크에 기초하여 제1 벡터를 획득할 수

있다.

[75] 여기서, 프로세서(120)는 제2 사용자 음성을 벡터화하여 특징 벡터를 획득할 수 있다. 프로세서(120)는 Mel-filter bank 나 MFCC(Mel-Frequency Cepstral Coefficients), Spectrogram 등을 이용하여 사용자 음성에 대응되는 특징 벡터를 획득할 수 있다.

[76] 그리고, 프로세서(120)는 제2 사용자 음성에 대응되는 특징 벡터를 제1 서브 네트워크에 입력하여 제1 벡터를 획득할 수 있다. 제1 서브 네트워크는 특징 벡터를 히든 벡터로 변환하는 네트워크를 의미할 수 있다. 제1 벡터는 히든 벡터를 의미할 수 있다.

[77] 제1 벡터와 관련된 구체적인 설명은 도 3의 수학식(211-1)을 통해 기재한다. 특징 벡터는 도3의 "X_t"에 해당할 수 있다. 또한, 제1 서브 네트워크는 도3의 "f_trans"에 대응될 수 있다. 또한, 제1 벡터는 도3의 "h_trans,t"에 대응될 수 있다.

[78] 한편, 프로세서(120)는 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득하고, 제1 인식 정보에 대응되는 원-핫 벡터, 제1 가중치 정보 및 제2 네트워크(220)에 포함된 제2 서브 네트워크에 기초하여 제2 벡터를 획득할 수 있다.

[79] 여기서, 프로세서(120)는 제1 사용자 음성(이전 사용자 음성)에 대응되는 제1 인식 정보를 획득하고, 제1 인식 정보에 대응되는 원-핫 벡터를 획득할 수 있다. 원-핫 벡터는 0 및 1로 이루어진 벡터를 의미할 수 있다. 또한, 원-핫 벡터는 벡터의 합이 1일 수 있다. 따라서, 원-핫 벡터는 '0'값을 가지는 복수의 벡터와 '1'값을 가지는 1개의 벡터를 포함할 수 있다.

[80] 여기서, 제2 네트워크(220)는 제1 가중치 정보를 포함할 수 있다. 제1 가중치 정보는 입력 임베딩(input embedding)일 수 있다. 제2 네트워크(220)는 제2 서브 네트워크를 포함할 수 있다. 제2 서브 네트워크는 제1 인식 정보에 대응되는 중간 벡터(또는 임베딩 벡터)를 히든 벡터로 변환하는 네트워크를 의미할 수 있다. 프로세서(120)는 제2 네트워크(220)를 통해 원-핫 벡터, 제1 가중치 정보 및 제2 서브 네트워크에 기초하여 제2 벡터를 획득할 수 있다.

[81] 여기서, 제2 벡터와 관련된 구체적인 설명은 도 3의 수학식(221-1) 및 수학식(222-1)을 통해 기재한다. 제1 인식 정보에 대응되는 원-핫 벡터는 도3의 "y_u-1"에 대응될 수 있다. 또한, 제1 가중치 정보는 도3의 "W_pred"에 대응될 수 있다. 또한, 중간 벡터는 도3의 "e_u-1"에 대응될 수 있다. 또한, 제2 서브 네트워크는 도3의 "f_pred"에 대응될 수 있다. 또한, 제2 벡터는 도3의 "h_pred,u"에 대응될 수 있다.

[82] 한편, 프로세서(120)는 제1 벡터, 제2 벡터 및 제3 네트워크(230)에 포함된 제3 서브 네트워크에 기초하여 제3 벡터를 획득하고, 제3 벡터 및 제2 가중치 정보에 기초하여 제2 인식 정보를 획득할 수 있다.

[83] 여기서, 제3 네트워크(230)는 제3 서브 네트워크를 포함할 수 있다. 제3 서브 네트워크는 제1 네트워크(210)에서 획득된 제1 벡터와 제2 네트워크(220)에서

획득된 제2 벡터를 결합하여 제3 벡터를 획득하는 네트워크일 수 있다. 제3 벡터는 히든 벡터일 수 있다.

[84] 여기서, 제3 네트워크(230)는 제2 가중치 정보를 포함할 수 있다. 제2 가중치 정보는 출력 임베딩(output embedding)일 수 있다. 프로세서(120)는 제2 가중치 정보 및 제3 벡터에 기초하여 제2 인식 정보를 획득할 수 있다. 구체적으로, 프로세서(120)는 제2 가중치 정보 및 제3 벡터를 곱셈하고, 곱셈된 값을 소프트맥스(softmax) 함수를 이용하여 정규화할 수 있다. 그리고, 프로세서(120)는 정규화된 값에 기초하여 제2 사용자 음성에 대응되는 제2 인식 정보를 획득할 수 있다.

[85] 여기서, 제2 인식 정보를 획득하는 동작은 도3의 수학식(231) 및 수학식(232)을 통해 기재한다. 제3 서브 네트워크는 도3의 "f_joint"에 대응될 수 있다. 제2 가중치 정보는 도3의 "W_joint"에 대응될 수 있다. 또한, 제3 벡터는 도3의 "h_joint"에 대응될 수 있다.

[86] 한편, 제1 가중치 정보는 기 설정된 개수의 서브워드에 대응되는 가중치를 포함하고, 제2 가중치 정보는 제1 가중치 정보에 포함된 가중치 및 추가 가중치를 포함할 수 있다.

[87] 여기서, 서브워드는 사용자가 발화한 음성으로 추측되는 기 설정된 단어를 의미할 수 있다. 서브워드는 음성 인식 모델에 따라 상이할 수 있다.

[88] 여기서, 제1 가중치 정보는 기 설정된 개수(V)의 서브워드에 대응되는 V개의 가중치를 포함할 수 있다. 여기서, V개의 가중치는 학습 동작에 의하여 결정될 수 있다. 제2 가중치 정보는 제1 가중치 정보에 포함된 V개의 가중치 및 추가 가중치를 더 포함할 수 있다.

[89] 여기서, 제1 가중치 정보에 포함된 가중치는 도8의 가중치(W_p1, W_p 2, W_p3, ... , W_pV)에 대응될 수 있다. 또한, 제2 가중치 정보에 포함된 추가 가중치는 도8의 추가 가중치(W_null)에 대응될 수 있다.

[90] 여기서, 프로세서(120)는 제1 가중치 정보에 포함된 가중치를 메모리(110)의 제1 영역에 저장하고, 추가 가중치를 메모리(110)의 제2 영역에 저장할 수 있다. 그리고, 프로세서(120)는 메모리(110)의 제1 영역에 저장된 가중치를 제1 가중치 정보로서 이용할 수 있다. 또한, 프로세서(120)는 메모리(110)의 제1 영역에 저장된 가중치 및 메모리(110)의 제2 영역에 저장된 추가 가중치를 제2 가중치 정보로서 이용할 수 있다. 메모리(110)에 가중치가 저장되는 구체적인 동작은 도5 내지 도 8에 기재한다.

[91] 한편, 추가 가중치는 제2 사용자 음성에 대응되는 서브워드가 존재하지 않는 경우 이용되는 가중치이고, 기 설정된 개수의 가중치의 차원(dimension)은 추가 가중치의 차원(dimension)과 동일할 수 있다.

[92] 여기서, 추가 가중치(W_null)는 사용자 음성이 V개의 서브워드 중 어느 것에도 해당하지 않는 경우 적용되는 가중치를 의미할 수 있다.

[93] 프로세서(120)는 음성 인식 모델(200)을 이용하여 사용자 음성이 수신되면

사용자 음성이 기 설정된 V개의 서브워드 각각에 대하여 얼마나 유사한지 판단할 수 있다. 예를 들어, 프로세서(120)는 사용자 음성이 제1 서브워드에 대응될 확률이 p1, 제2 서브워드에 대응될 확률이 p2, ... , V번째 서브워드에 대응될 확률이 pV라고 판단할 수 있다. 그리고, 프로세서(120)는 p1 내지 pV 중 가장 높은 확률값을 갖는 서브워드를 사용자 음성에 대응되는 인식 정보로 결정할 수 있다.

[94] 여기서, 프로세서(120)는 p1 내지 pV 중 가장 높은 확률 값이 임계값 이상인지 추가로 확인할 수 있다. 가장 높은 확률값이 임계값 이상이면, 프로세서(120)는 가장 높은 확률값을 갖는 서브워드가 사용자 음성에 대응되는 인식 정보로 결정할 수 있다.

[95] 한편, 가장 높은 확률값이 임계값 미만이면, 프로세서(120)는 사용자 음성에 대응되는 서브워드가 존재하지 않는 것으로 판단할 수 있다. 여기서, 프로세서(120)는 사용자 음성에 대응되는 서브워드가 존재하지 않으면, 추가 가중치(W_null)를 이용하여 사용자 음성에 대응되는 인식 정보를 획득할 수 있다.

[96] 여기서, 제1 가중치 정보 및 제2 가중치 정보에 포함된 가중치의 차원은 동일할 수 있다. 가중치의 차원과 관련된 표현은 도 6 및 도 8에서 기재한다.

[97] 한편, 제1 가중치 정보는 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트(gradient), 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트(gradient) 및 학습률(Learning rate)에 기초하여 학습되고, 제2 가중치 정보는 학습된 제1 가중치 정보에 기초하여 결정될 수 있다.

[98] 여기서, 프로세서(120)는 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트(gradient) 및 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트(gradient)를 획득할 수 있다. 그리고, 프로세서(120)는 제1 그래디언트, 제2 그래디언트 및 학습률(Learning rate)에 기초하여 제1 가중치 정보를 학습할 수 있다. 그리고, 프로세서(120)는 학습된 제1 가중치 정보에 기초하여 제2 가중치 정보를 결정할 수 있다.

[99] 이와 관련된 구체적인 동작은 도 14 및 도 15에서 기재한다. 여기서, 제1 그래디언트는 도 14의 " $\nabla_{W\_pred}L$"에 대응될 수 있다. 또한, 제2 그래디언트는 도 14의 " $\nabla_{W\_joint}L$"에 대응될 수 있다. 또한, 학습률은 도 14의 η(에크)에 대응될 수 있다. 또한, 학습 동작이 수행되기 전의 제1 가중치 정보는 도 14의 "W_pred-old"에 대응될 수 있다. 또한, 학습 동작이 수행된 후의 제1 가중치 정보는 도 14의 "W_pred-new"에 대응될 수 있다. 또한, 학습 동작이 수행되기 전의 제2 가중치 정보는 도 14의 "W_joint-old"에 대응될 수 있다. 또한, 학습 동작이 수행된 후의 제2 가중치 정보는 도 14의 "W_joint-new"에 대응될 수 있다.

[100] 한편, 제1 가중치 정보 및 제2 가중치 정보는 제1 서브 가중치 정보 및 제2 서브 가중치 정보의 평균값에 기초하여 학습되고, 제1 서브 가중치는 제1 가중치

정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트 및 학습률에 기초하여 산출되고, 제2 서브 가중치는 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트 및 학습률에 기초하여 산출될 수 있다.

[101] 여기서, 프로세서(120)는 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트 및 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트를 획득할 수 있다. 그리고, 프로세서(120)는 제1 그래디언트 및 학습률에 기초하여 제1 서브 가중치 정보를 획득하고, 제2 그래디언트 및 학습률에 기초하여 제2 서브 가중치 정보를 획득할 수 있다. 그리고, 프로세서(120)는 제1 서브 가중치 정보 및 제2 서브 가중치 정보의 평균값에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습할 수 있다.

[102] 이와 관련된 구체적인 동작은 도 16 및 도 17에서 기재한다. 여기서, 제1 그래디언트는 도 16의 " $\nabla_{W\_pred}L$"에 대응될 수 있다. 또한, 제2 그래디언트는 도 16의 " $\nabla_{W\_joint}L$"에 대응될 수 있다. 또한, 학습률은 도 16의 η(에크)에 대응될 수 있다. 또한, 제1 서브 가중치 정보는 도 16의 "W_pred-sub"에 대응될 수 있다. 또한, 제2 서브 가중치 정보는 도 16의 "W_joint-sub"에 대응될 수 있다. 또한, 학습 동작이 수행되기 전의 제1 가중치 정보는 도 16의 "W_pred-old"에 대응될 수 있다. 또한, 학습 동작이 수행된 후의 제1 가중치 정보는 도 16의 "W_pred-new"에 대응될 수 있다. 또한, 학습 동작이 수행되기 전의 제2 가중치 정보는 도 16의 "W_joint-old"에 대응될 수 있다. 또한, 학습 동작이 수행된 후의 제2 가중치 정보는 도 16의 "W_joint-new"에 대응될 수 있다.

[103] 한편, 전자 장치(100)는 마이크(미도시)를 더 포함할 수 있다.

[104] 여기서, 마이크(미도시)는 사용자 음성이나 기타 소리를 입력 받아 오디오 데이터로 변환하기 위한 구성이다. 마이크(미도시)는 활성화 상태에서 사용자의 음성을 수신할 수 있다. 예를 들어, 마이크(미도시)는 전자 장치(100)의 상측이나 전면 방향, 측면 방향 등에 일체형으로 형성될 수 있다. 마이크(미도시)는 아날로그 형태의 사용자 음성을 수집하는 마이크, 수집된 사용자 음성을 증폭하는 앰프 회로, 증폭된 사용자 음성을 샘플링하여 디지털 신호로 변환하는 A/D 변환회로, 변환된 디지털 신호로부터 노이즈 성분을 제거하는 필터 회로 등과 같은 다양한 구성을 포함할 수 있다.

[105] 여기서, 프로세서(120)는 마이크(미도시)를 통해 사용자 음성을 획득할 수 있다. 그리고, 프로세서(120)는 사용자 음성에 대응되는 인식 정보를 음성 인식 모델(200)로부터 획득할 수 있다.

[106] 전자 장치(100)는 제2 네트워크(220)에 포함된 제1 가중치 정보와 제3 네트워크(230)에 포함된 제2 가중치 정보를 이용하여 사용자 음성에 대응되는 인식 정보를 획득할 수 있다. 여기서, 제1 가중치 정보에 포함된 가중치가 제2 가중치 정보에도 포함될 수 있다. 제2 가중치 정보는 제1 가중치 정보에 포함된 가중치를 그대로 이용할 수 있으므로, 일부 가중치를 공유할 수 있다. 따라서,

전자 장치(100)는 음성 인식 성능의 하락 없이 모델의 가중치가 차지하는 크기를 감소시킬 수 있다.

[107] 한편, 이상에서는 전자 장치(100)를 구성하는 간단한 구성에 대해서만 도시하고 설명하였지만, 구현 시에는 다양한 구성이 추가로 구비될 수 있다.

[108] 도 2는 복수의 네트워크로 구성된 음성 인식 모델(200)을 설명하기 위한 도면이다.

[109] 도 2를 참조하면, 음성 인식 모델(200)은 제1 네트워크(210), 제2 네트워크(220) 및 제3 네트워크(230)를 포함할 수 있다.

[110] 여기서, 음성 인식 모델(200)은 인공 지능 모델일 수 있다. 예를 들어, 음성 인식 모델(200)은 RNN-T(Recurrent Neural Network Transducer) 모델일 수 있다.

[111] 여기서, 제1 네트워크(210)는 전사 네트워크(Transcription Network)를 의미할 수 있다. 여기서, 제1 네트워크(210)는 사용자 음성을 입력 받아 사용자 음성에 대응되는 제1 벡터를 획득할 수 있다.

[112] 여기서, 제2 네트워크(220)는 예측 네트워크(Prediction Network)를 의미할 수 있다. 여기서, 제2 네트워크(220)는 이전 출력 결과를 입력 받아 제2 벡터를 획득할 수 있다. 여기서, 이전 출력 결과는 이전 사용자 음성에 대응되는 인식 정보를 의미할 수 있다. 여기서, 제2 네트워크(220)는 제1 가중치 정보를 이용하여 이전 출력 결과에 대응되는 제2 벡터를 획득할 수 있다.

[113] 여기서, 제3 네트워크(230)는 조인트 네트워크(Joint Network)를 의미할 수 있다. 여기서, 제3 네트워크(230)는 제1 네트워크(210)에서 획득된 제1 벡터 및 제2 네트워크(220)에서 획득된 제2 벡터를 입력 받아 사용자 음성에 대응되는 출력 결과를 획득할 수 있다. 여기서, 출력 결과는 사용자 음성에 대응되는 타겟 워드를 의미할 수 있다.

[114] 도 3은 이전 출력값에 기초하여 인식 정보를 획득하는 음성 인식 모델(200)을 설명하기 위한 도면이다.

[115] 도 3을 참조하면, 음성 인식 모델(200)은 제1 네트워크(210), 제2 네트워크(220) 및 제3 네트워크(230)를 포함할 수 있다.

[116] 여기서, 제1 네트워크(210)는 사용자 음성(X_t)을 획득할 수 있다. 그리고, 제1 네트워크(210)는 수학식(211-1)을 이용하여 제1 벡터(h_trans,t)를 획득할 수 있다. 구체적으로, 제1 네트워크(210)는 수신된 사용자 음성(X_t)을 제1 서브 네트워크(f_trans)에 입력하여 제1 벡터(h_trans,t)를 획득할 수 있다. 여기서, 제1 서브 네트워크(f_trans)는 제1 네트워크(210)에 포함된 네트워크일 수 있다.

[117] 또한, 제2 네트워크(220)는 이전 사용자 음성에 대응되는 인식 정보(y_u-1)를 획득할 수 있다. 그리고, 제2 네트워크(220)는 수학식(221-1) 및 수학식(222-1)을 이용하여 제2 벡터(h_pred,u)를 획득할 수 있다. 구체적으로, 수학식(221-1)을 통해, 제2 네트워크(220)는 제1 가중치 정보(W_pred) 및 이전 사용자 음성에 대응되는 인식 정보(y_u-1)를 곱셈하여 중간 벡터(e_u-1)를 획득할 수 있다. 여기서, 중간 벡터(e_u-1)는 임베딩 벡터를 의미할 수 있다. 그리고,

수학식(222-1)을 통해, 제2 네트워크(220)는 중간 벡터(e_u-1)를 제2 서브 네트워크(f_pred)에 입력하여 제2 벡터(h_pred,u)를 획득할 수 있다.

[118] 또한, 제3 네트워크(230)는 사용자 음성(X_t)에 대응되는 인식 정보(y_u)를 획득할 수 있다. 여기서, 사용자 음성(X_t)에 대응되는 인식 정보(y_u)는 사용자 음성에 대응되는 인식 정보(또는 타겟 워드)를 의미할 수 있다. 그리고, 제3 네트워크(230)는 수학식(231) 및 수학식(232)을 이용하여 사용자 음성에 대응되는 인식 정보를 획득할 수 있다. 구체적으로, 수학식(231)을 통해, 제3 네트워크(230)는 제1 벡터(h_trans,t) 및 제2 벡터(h_pred,u)를 제3 서브 네트워크(f_joint)에 입력하여 제3 벡터(h_joint)를 획득할 수 있다. 그리고, 수학식(232)을 통해, 제3 네트워크(230)는 제2 가중치 정보(W_joint) 및 제3 벡터(h_joint)를 소프트맥스(Softmax)에 입력하여 사용자 음성에 대응되는 인식 정보(y_u)를 획득할 수 있다. 수학식(232)에서 p(y_ulX_t,y_u-1)는 사용자 음성(X_t) 및 이전 사용자 음성에 대응되는 인식 정보(y_u-1)에 기초하여 결정된 인식 정보(y_u)에 대한 확률 값을 의미할 수 있다.

[119] 한편, 도 3에서는 이전 사용자 음성에 대응되는 인식 정보가 1개인 경우를 가정하여 설명하였다. 하지만, 이전 사용자 음성에 대응되는 인식 정보가 복수 개일 수 있으며 이와 관련된 설명은 도 4에서 기재한다.

[120] 도 4는 복수의 이전 출력값들에 기초하여 인식 정보를 획득하는 음성 인식 모델(200)을 설명하기 위한 도면이다.

[121] 도 4를 참조하면, 음성 인식 모델(200)은 제1 네트워크(210), 제2 네트워크(220) 및 제3 네트워크(230)를 포함할 수 있다.

[122] 여기서, 제1 네트워크(210)는 사용자 음성(X_1:t)을 획득할 수 있다. 그리고, 제1 네트워크(210)는 수학식(211-2)을 이용하여 제1 벡터(h_trans,t)를 획득할 수 있다. 구체적으로, 제1 네트워크(210)는 수신된 사용자 음성(X_1:t)을 제1 서브 네트워크(f_trans)에 입력하여 제1 벡터(h_trans,t)를 획득할 수 있다. 여기서, 제1 서브 네트워크(f_trans)는 제1 네트워크(210)에 포함된 네트워크일 수 있다.

[123] 또한, 제2 네트워크(220)는 이전 사용자 음성에 대응되는 인식 정보(y_1:u-1)를 획득할 수 있다. 그리고, 제2 네트워크(220)는 수학식(221-2) 및 수학식(222-2)을 이용하여 제2 벡터(h_pred,u)를 획득할 수 있다. 구체적으로, 수학식(221-2)을 통해, 제2 네트워크(220)는 제1 가중치 정보(W_pred) 및 이전 사용자 음성에 대응되는 인식 정보(y_1:u-1)를 곱셈하여 중간 벡터(e_1:u-1)를 획득할 수 있다. 여기서, 중간 벡터(e_1:u-1)는 임베딩 벡터를 의미할 수 있다. 그리고, 수학식(222-2)을 통해, 제2 네트워크(220)는 중간 벡터(e_1:u-1)를 제2 서브 네트워크(f_pred)에 입력하여 제2 벡터(h_pred,u)를 획득할 수 있다.

[124] 또한, 제3 네트워크(230)는 사용자 음성(X_1:t)에 대응되는 인식 정보(y_u)를 획득할 수 있다. 여기서, 사용자 음성(X_1:t)에 대응되는 인식 정보(y_u)는 사용자 음성에 대응되는 인식 정보(또는 타겟 워드)를 의미할 수 있다. 그리고, 제3 네트워크(230)는 수학식(231) 및 수학식(232)을 이용하여 사용자 음성에

대응되는 인식 정보를 획득할 수 있다. 구체적으로, 수학식(231)을 통해, 제3 네트워크(230)는 제1 벡터(h_trans,t) 및 제2 벡터(h_pred,u)를 제3 서브 네트워크(f_joint)에 입력하여 제3 벡터(h_joint)를 획득할 수 있다. 그리고, 수학식(232)을 통해, 제3 네트워크(230)는 제2 가중치 정보(W_joint) 및 제3 벡터(h_joint)를 소프트맥스(Softmax)에 입력하여 사용자 음성에 대응되는 인식 정보(y_u)를 획득할 수 있다. 수학식(232)에서 p(y_ulX_1:t,y_1:u-1)는 사용자 음성(X_1:t) 및 이전 사용자 음성에 대응되는 인식 정보(y_1:u-1)에 기초하여 결정된 인식 정보(y_u)에 대한 확률 값을 의미할 수 있다.

[125] 도 5는 제1 가중치 정보 및 제2 가중치 정보가 동일하지 않은 실시 예의 가중치 정보 저장 방식을 설명하기 위한 도면이다.

[126] 도 5를 참조하면, 음성 인식 모델(200)은 제1 네트워크(210), 제2 네트워크(220) 및 제3 네트워크(230)를 포함할 수 있다. 여기서, 제2 네트워크(220)는 제1 가중치 정보(W_pred)를 이용하는 네트워크일 수 있다. 여기서, 제3 네트워크(230)는 제2 가중치 정보(W_joint)를 이용하는 네트워크일 수 있다. 여기서, 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)는 전자 장치(100)의 메모리(110)에 저장될 수 있다.

[127] 여기서, 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)는 상이한 가중치를 포함할 수 있다. 따라서, 전자 장치(100)는 메모리(110)의 제1 영역(510)에 제1 가중치 정보(W_pred)를 저장하고, 메모리(110)의 제2 영역(520)에 제2 가중치 정보(W_joint)를 저장할 수 있다.

[128] 도 6는 제1 가중치 정보 및 제2 가중치 정보가 동일하지 않은 실시 예의 가중치 정보 구성을 설명하기 위한 도면이다.

[129] 도 6을 참조하면, 제1 가중치 정보(W_pred, 610) 및 제2 가중치 정보(W_joint, 620)는 서로 다른 가중치를 포함할 수 있다.

[130] 여기서, 제1 가중치 정보(W_pred, 610)는 D차원의 V개의 가중치를 포함할 수 있다. 여기서, V는 기 설정된 서브워드의 개수를 의미할 수 있다. 여기서, D는 기 설정된 서브워드 가중치의 차원(dimension)을 의미할 수 있다. 여기서, 제1 가중치 정보(W_pred, 610)는 제1 서브워드 가중치(W_p1), 제2 서브워드 가중치(W_p2), 제3 서브워드 가중치(W_p3) 내지 V번째 서브워드 가중치(W_pV)를 포함할 수 있다.

[131] 여기서, 제2 가중치 정보(W_joint, 620)는 D차원의 V+1개의 가중치를 포함할 수 있다. 여기서, V는 기 설정된 서브워드의 개수를 의미할 수 있다. 여기서, D는 기 설정된 서브워드 가중치의 차원(dimension)을 의미할 수 있다. 여기서, 제2 가중치 정보(W_joint, 620)에 포함된 가중치는 제1 서브워드 가중치(W_j1), 제2 서브워드 가중치(W_j2), 제3 서브워드 가중치(W_j3) 내지 V번째 서브워드 가중치(W_jV) 및 추가 가중치(W_null)를 포함할 수 있다. 여기서, 추가 가중치(W_null)는 사용자 음성이 V개의 서브워드 중 어느 것에도 해당하지 않는 경우 적용되는 가중치를 의미할 수 있다. 따라서, 제2 가중치 정보(W_joint,

620)는 V개의 서브워드에 대응되는 가중치(W_j1, W_j2, W_j3, ... , W_jV) 및 추가 가중치(W_null)를 포함할 수 있다. 여기서, 제2 가중치 정보(W_joint, 620)는 총 V+1개의 가중치를 포함할 수 있다.

[132] 도 7은 제1 가중치 정보 및 제2 가중치 정보가 일부 동일한 실시 예의 가중치 정보 저장 방식을 설명하기 위한 도면이다.

[133] 도 7을 참조하면, 음성 인식 모델(200)은 제1 네트워크(210), 제2 네트워크(220) 및 제3 네트워크(230)를 포함할 수 있다. 여기서, 제2 네트워크(220)는 제1 가중치 정보(W_pred)를 이용하는 네트워크일 수 있다. 여기서, 제3 네트워크(230)는 제2 가중치 정보(W_joint)를 이용하는 네트워크일 수 있다. 여기서, 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)는 전자 장치(100)의 메모리(110)에 저장될 수 있다.

[134] 여기서, 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)는 동일한 가중치를 포함할 수 있다. 구체적으로, 제1 가중치 정보(W_pred)가 포함하는 가중치를 제2 네트워크(220)도 포함할 수 있다. 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)는 공통의 가중치를 공유할 수 있다. 따라서, 전자 장치(100)는 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)를 별도로 저장하지 않아도 된다. 전자 장치(100)는 제1 가중치 정보(W_pred)에 포함된 가중치를 그대로 제2 가중치 정보(W_joint)로서 이용할 수 있기 때문이다.

[135] 구체적으로, 전자 장치(100)는 메모리(110)의 제1 영역(710)에 제1 가중치 정보(W_pred)에 포함된 가중치를 저장할 수 있다. 그리고, 전자 장치(100)는 메모리(110)의 제2 영역(720)에 추가 가중치(W_null)를 저장할 수 있다.

[136] 여기서, 전자 장치(100)는 메모리(110)의 제1 영역(710)에 저장된 가중치를 제1 가중치 정보(W_pred)로서 이용할 수 있다. 그리고, 전자 장치(100)는 메모리(110)의 제1 영역(710)에 저장된 가중치 및 메모리(110)의 제2 영역(720)에 저장된 가중치(W_null)를 제2 가중치 정보(W_joint)로서 이용할 수 있다. 결과적으로, 도 7의 실시 예는 도 5의 실시 예보다 메모리(110)의 저장 공간을 줄일 수 있다.

[137] 도 8은 제1 가중치 정보 및 제2 가중치 정보가 일부 동일한 실시 예의 가중치 정보 구성을 설명하기 위한 도면이다.

[138] 도 8을 참조하면, 제1 가중치 정보(W_pred, 810) 및 제2 가중치 정보(W_joint-new, 820)는 서로 동일한 가중치를 포함할 수 있다.

[139] 여기서, 제1 가중치 정보(W_pred, 810)는 D차원의 V개의 가중치를 포함할 수 있다. 여기서, V는 기 설정된 서브워드의 개수를 의미할 수 있다. 여기서, D는 기 설정된 서브워드 가중치의 차원(dimension)을 의미할 수 있다. 여기서, 제1 가중치 정보(W_pred, 810)는 제1 서브워드 가중치(W_p1), 제2 서브워드 가중치(W_p2), 제3 서브워드 가중치(W_p3) 내지 V번째 서브워드 가중치(W_pV)를 포함할 수 있다.

[140] 여기서, 제2 가중치 정보(W_joint-new, 820)는 D차원의 V+1개의 가중치를

포함할 수 있다. 여기서, V는 기 설정된 서브워드의 개수를 의미할 수 있다. 여기서, D는 기 설정된 서브워드 가중치의 차원(dimension)을 의미할 수 있다.

[141] 여기서, 제2 가중치 정보(W_joint-new, 820)는 제1 가중치 정보(W_pred, 810)에 포함된 가중치 및 추가 가중치(W_null)를 포함할 수 있다. 구체적으로, 제2 가중치 정보(W_joint-new, 820)에 포함된 가중치는 제1 서브워드 가중치(W_p1), 제2 서브워드 가중치(W_p2), 제3 서브워드 가중치(W_p3) 내지 V번째 서브워드 가중치(W_pV) 및 추가 가중치(W_null)를 포함할 수 있다. 여기서, 추가 가중치(W_null)는 사용자 음성이 V개의 서브워드 중 어느 것에도 해당하지 않는 경우 적용되는 가중치를 의미할 수 있다. 따라서, 제2 가중치 정보(W_joint-new, 820)는 V개의 서브워드에 대응되는 가중치(W_p1, W_p 2, W_p3, ... , W_pV) 및 추가 가중치(W_null)를 포함할 수 있다. 여기서, 제2 가중치 정보(W_joint-new, 820)는 총 V+1개의 가중치를 포함할 수 있다.

[142] 여기서, 전자 장치(100)는 수학식(830)에 기초하여 제2 가중치 정보(W_joint-new, 820)를 획득할 수 있다. 구체적으로, 전자 장치(100)는 제1 가중치 정보(W_pred, 810)의 전치 행렬(transposed matrix)을 획득할 수 있다. 그리고, 전자 장치(100)는 제1 가중치 정보(W_pred, 810)의 전치 행렬(transposed matrix)에 추가 가중치(W_null)를 추가하여 제2 가중치 정보(W_joint-new, 830)를 획득할 수 있다.

[143] 도 9는 음성 인식 모델(200)을 이용하여 사용자 음성에 대응되는 인식 정보를 획득하는 동작을 설명하기 위한 흐름도이다.

[144] 도 9를 참조하면, 전자 장치(100)는 사용자 음성을 수신할 수 있다 (S905). 그리고, 전자 장치(100)는 사용자 음성을 복수의 네트워크로 구성된 음성 인식 모델(200)에 입력할 수 있다 (S910). 그리고, 전자 장치(100)는 음성 인식 모델(200)로부터 사용자 음성에 대응되는 인식 정보를 획득할 수 있다 (S915). 여기서, 사용자 음성에 대응되는 인식 정보는 음성 인식 모델(200)로부터 출력될 수 있다. 사용자 음성은 음성 인식 모델(200)에 입력되는 입력 데이터이고, 사용자 음성에 대응되는 인식 정보는 음성 인식 모델(200)로부터 출력되는 출력 데이터일 수 있다.

[145] 도 10은 제1 사용자 음성 및 제2 사용자 음성에 기초하여 인식 정보를 획득하는 동작을 설명하기 위한 흐름도이다.

[146] 도 10을 참조하면, 전자 장치(100)는 제1 사용자 음성에 대응되는 제1 인식 정보를 저장할 수 있다 (S1010). 여기서, 전자 장치(100)는 제1 사용자 음성을 음성 인식 모델(200)에 입력하여 제1 사용자 음성에 대응되는 제1 인식 정보를 출력 데이터로서 획득할 수 있다. 그리고, 전자 장치(100)는 제1 인식 정보를 메모리(110)에 저장할 수 있다.

[147] 여기서, 전자 장치(100)는 제2 사용자 음성을 수신할 수 있다 (S1020). 그리고, 전자 장치(100)는 제2 사용자 음성을 제1 네트워크(210)에 입력하여 제1 벡터를 획득할 수 있다 (S1030). 제1 벡터를 획득하는 동작은 제1 네트워크(210)에서

수행될 수 있다.

[148] 그리고, 전자 장치(100)는 제1 인식 정보를 제2 네트워크(220)에 입력하여 제2 벡터를 획득할 수 있다 (S1040). 제2 벡터를 획득하는 동작은 제2 네트워크(220)에서 수행될 수 있다.

[149] 그리고, 전자 장치(100)는 제1 벡터 및 제2 벡터를 제3 네트워크(230)에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보를 획득할 수 있다 (S1050). 제2 인식 정보를 획득하는 동작은 제3 네트워크(230)에서 수행될 수 있다.

[150] 도 11은 제1 벡터를 획득하는 구체적인 동작을 설명하기 위한 흐름도이다.

[151] 도 11을 참조하면, S1110, S1120, S1140, S1150 단계는 도 10의 S1010, S1020, S1040, S1050 단계에 대응될 수 있는 바 중복 설명을 생략한다.

[152] 여기서, 전자 장치(100)는 제2 사용자 음성을 수신하는 단계 (S1120) 이후, 제2 사용자 음성에 대응되는 특징 벡터를 획득할 수 있다 (S1131). 여기서, 특징 벡터는 사용자 음성에 기초하여 획득된 벡터를 의미할 수 있다. 그리고, 전자 장치(100)는 제2 사용자 음성에 대응되는 특징 벡터 및 제1 네트워크(210)에 포함된 제1 서브 네트워크(f_trans)에 기초하여 제1 벡터(h_trans,t)를 획득할 수 있다 (S1132).

[153] 도 12는 제2 벡터를 획득하는 구체적인 동작을 설명하기 위한 흐름도이다.

[154] 도 12를 참조하면, S1210, S1220, S1230, S1250 단계는 도 10의 S1010, S1020, S1030, S1050 단계에 대응될 수 있는 바 중복 설명을 생략한다.

[155] 여기서, 전자 장치(100)는 제1 벡터를 획득하는 단계 (S1230) 이후, 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득할 수 있다 (S1241). 여기서, 원-핫 벡터는 0 및 1로 이루어진 벡터를 의미할 수 있다. 또한, 원-핫 벡터는 벡터의 합이 1일 수 있다. 따라서, 원-핫 벡터는 '0'값을 가지는 복수의 벡터와 '1'값을 가지는 1개의 벡터를 포함할 수 있다.

[156] 그리고, 전자 장치(100)는 제1 인식 정보에 대응되는 원-핫 벡터, 제1 가중치 정보(W_pred) 및 제2 네트워크(220)에 포함된 제2 서브 네트워크(f_pred)에 기초하여 제2 벡터(h_pred,u)를 획득할 수 있다 (S1242).

[157] 도 13은 제3 벡터를 획득하는 구체적인 동작을 설명하기 위한 흐름도이다.

[158] 도 13을 참조하면, S1310, S1320, S1330, S1340 단계는 도 10의 S1010, S1020, S1030, S1040 단계에 대응될 수 있는 바 중복 설명을 생략한다.

[159] 여기서, 전자 장치(100)는 제2 벡터를 획득하는 단계 (S1340) 이후, 제1 벡터, 제2 벡터 및 제3 네트워크(230)에 포함된 제3 서브 네트워크(f_joint)에 기초하여 제3 벡터(h_joint)를 획득할 수 있다 (S1352).

[160] 도 14는 일 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 도면이다.

[161] 도 14를 참조하면, 일 실시 예에 따라, 전자 장치(100)는 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)를 학습할 수 있다.

[162] 전자 장치(100)는 제1 가중치 정보(W_pred)에 따른 손실값의 변화량을

나타내는 제1 그래디언트($\nabla_{W\_pred}L$)를 획득할 수 있다 (S1410-1). 또한, 전자 장치(100)는 제2 가중치 정보(W_joint)에 따른 손실값의 변화량을 나타내는 제2 그래디언트($\nabla_{W\_joint}L$)를 획득할 수 있다 (S1410-2).

[163] 여기서, L은 손실 함수(Loss Function)에 기초하여 획득되는 손실값을 의미할 수 있다.

[164] 여기서, 그래디언트는 기울기 벡터를 의미할 수 있다. 구체적으로, 제1 그래디언트($\nabla_{W\_pred}L$)는 제1 가중치 정보(W_pred)가 변함에 따라 손실값이 얼마나 변하는지 나타내는 기울기 벡터를 의미할 수 있다. 또한, 제2 그래디언트($\nabla_{W\_joint}L$)는 제2 가중치 정보(W_joint)가 변함에 따라 손실값이 얼마나 변하는지 나타내는 기울기 벡터를 의미할 수 있다.

[165] 그리고, 전자 장치(100)는 업데이트된 제1 가중치 정보(W_pred-new)를 획득할 수 있다 (S1420). 전자 장치(100)는 제1 가중치 정보(W_pred-old), 학습률(에크, η), S1410-1 단계에서 획득한 제1 그래디언트($\nabla_{W\_pred}L$) 및 S1410-2 단계에서 획득한 제2 그래디언트($\nabla_{W\_joint}L$)에 기초하여 업데이트된 제1 가중치 정보(W_pred-new)를 획득할 수 있다. 구체적으로, 전자 장치(100)는 제1 그래디언트($\nabla_{W\_pred}L$) 및 제2 그래디언트($\nabla_{W\_joint}L$)의 합산값($\nabla_{W\_pred}L+\nabla_{W\_joint}L$)을 획득하고, 획득된 합산값에 학습률(에크, η)을 곱하여 중간값($\eta(\nabla_{W\_pred}L+\nabla_{W\_joint}L)$)을 획득할 수 있다. 그리고, 전자 장치(100)는 제1 가중치 정보(W_pred-old)에서 중간값($\eta(\nabla_{W\_pred}L+\nabla_{W\_joint}L)$)을 뺄셈하여 업데이트된 제1 가중치 정보(W_pred-new)를 획득할 수 있다.

[166] 그리고, 전자 장치(100)는 업데이트된 제2 가중치 정보(W_joint-new)를 획득할 수 있다 (S1430). 전자 장치(100)는 제2 가중치 정보(W_joint-old)에 업데이트된 제1 가중치 정보(W_pred-new)를 대입하여 업데이트된 제2 가중치 정보(W_joint-new)를 획득할 수 있다. 여기서, S1430 단계는 도 8의 수학식(830)에 대응될 수 있다.

[167] 여기서, 업데이트된 제1 가중치 정보(W_pred-new) 및 업데이트된 제2 가중치 정보(W_joint-new)는 동일한 가중치를 포함할 수 있다. 여기서, 업데이트된 제2 가중치 정보(W_joint-new)는 제1 가중치 정보(W_pred-new)보다 추가 가중치(W_null)를 더 포함할 수 있다.

[168] 도 15는 일 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 흐름도이다.

[169] 도 15를 참조하면, 전자 장치(100)는 제1 가중치 정보(W_pred)에 따른 손실값의 변화량을 나타내는 제1 그래디언트($\nabla_{W\_pred}L$)를 획득하고, 제2 가중치 정보(W_joint)에 따른 손실값의 변화량을 나타내는 제2 그래디언트($\nabla_{W\_joint}L$)를 획득할 수 있다 (S1510). 여기서, S1510 단계는 도 14의 S1410-1, S1410-2 단계에 대응될 수 있다.

[170] 또한, 전자 장치(100)는 제1 그래디언트($\nabla_{W\_pred}L$) 및 제2 그래디언트($\nabla_{W\_joint}L$)가 합산된 값($\nabla_{W\_pred}L+\nabla_{W\_joint}L$)을 획득할 수 있다 (S1521).

[171] 그리고, 전자 장치(100)는 S1521 단계의 합산된 값($\nabla_{W\_pred}L+\nabla_{W\_joint}L$)에 학습률(에크, η)을 곱한 값(η($\nabla_{W\_pred}L+\nabla_{W\_joint}L$))을 획득할 수 있다 (S1522). 그리고, 전자 장치(100)는 제1 가중치 정보(W_pred-old) 및 S1522 단계에서 획득한 값(η($\nabla_{W\_pred}L+\nabla_{W\_joint}L$))에 기초하여 업데이트된 제1 가중치 정보(W_pred-new)를 획득할 수 있다 (S1523). 여기서, S1521, S1522, S1523 단계는 도 14의 S1420 단계에 대응될 수 있다.

[172] 또한, 전자 장치(100)는 업데이트된 제1 가중치 정보(W_pred-new)에 기초하여 업데이트된 제2 가중치 정보(W_joint-new)를 획득할 수 있다 (S1530). 구체적으로, 업데이트된 제2 가중치 정보(W_joint-new)는 업데이트된 제1 가중치 정보(W_pred-new)에 포함된 가중치 및 추가 가중치(W_null)를 포함할 수 있다. 여기서, S1530 단계는 도 14의 S1430 단계에 대응될 수 있다.

[173] 도 16은 다른 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 도면이다.

[174] 도 16을 참조하면, 다른 실시 예에 따라, 전자 장치(100)는 제1 가중치 정보(W_pred) 및 제2 가중치 정보(W_joint)를 학습할 수 있다.

[175] 여기서, 전자 장치(100)는 제1 가중치 정보(W_pred)에 따른 손실값의 변화량을 나타내는 제1 그래디언트($\nabla_{W\_pred}L$)를 획득할 수 있다 (S1610-1). 또한, 전자 장치(100)는 제2 가중치 정보(W_joint-old)에 따른 손실값의 변화량을 나타내는 제2 그래디언트($\nabla_{W\_joint}L$)를 획득할 수 있다 (S1610-2).

[176] 여기서, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub)를 획득할 수 있다 (S1620-1). 구체적으로, 전자 장치(100)는 학습률(에크, η)에 제1 그래디언트($\nabla_{W\_pred}L$)를 곱한 값(η $\nabla_{W\_pred}L$)을 획득할 수 있다. 그리고, 전자 장치(100)는 제1 가중치 정보(W_pred-old)에서 값(η $\nabla_{W\_pred}L$)을 뺄셈하여 제1 서브 가중치 정보(W_pred-sub)를 획득할 수 있다.

[177] 여기서, 전자 장치(100)는 제2 서브 가중치 정보(W_joint-sub)를 획득할 수 있다 (S1620-2). 구체적으로, 전자 장치(100)는 학습률(에크, η)에 제2 그래디언트($\nabla_{W\_joint}L$)를 곱한 값(η $\nabla_{W\_joint}L$)을 획득할 수 있다. 그리고, 전자 장치(100)는 제2 가중치 정보(W_joint-old)에서 값(η $\nabla_{W\_joint}L$)을 뺄셈하여 제2 서브 가중치 정보(W_joint-sub)를 획득할 수 있다.

[178] 여기서, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub) 및 제2 서브 가중치 정보(W_joint-sub)에 기초하여 업데이트된 제1 가중치 정보(W_pred-new)를 획득할 수 있다 (S1630-1). 구체적으로, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub) 및 제2 서브 가중치 정보(W_joint-sub)의 평균값을 업데이트된 제1 가중치 정보(W_pred-new)로서 획득할 수 있다.

[179] 여기서, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub) 및 제2 서브 가중치 정보(W_joint-sub)에 기초하여 업데이트된 제2 가중치 정보(W_joint-new)를 획득할 수 있다 (S1630-2). 구체적으로, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub) 및 제2 서브 가중치 정보(W_joint-sub)의 평균값을 업데이트된 제2 가중치 정보(W_joint-new)로서 획득할 수 있다.

[180] 여기서, 업데이트된 제1 가중치 정보(W_pred-new) 및 업데이트된 제2 가중치 정보(W_joint-new)는 동일한 가중치를 포함할 수 있다. 여기서, 업데이트된 제2 가중치 정보(W_joint-new)는 제1 가중치 정보(W_pred-new)보다 추가 가중치(W_null)를 더 포함할 수 있다.

[181] 도 17은 다른 실시 예에 따른 학습 방법에 기초하여 제1 가중치 정보 및 제2 가중치 정보를 학습하는 동작을 설명하기 위한 흐름도이다.

[182] 도 17을 참조하면, 전자 장치(100)는 제1 가중치 정보(W_pred)에 따른 손실값의 변화량을 나타내는 제1 그래디언트($\nabla_{W\_pred}L$)를 획득하고, 제2 가중치 정보(W_joint)에 따른 손실값의 변화량을 나타내는 제2 그래디언트($\nabla_{W\_joint}L$)를 획득할 수 있다 (S1710). 여기서, S1710 단계는 도 16의 S1610-1, S1610-2 단계에 대응될 수 있다.

[183] 또한, 전자 장치(100)는 학습률(에크, η)에 제1 그래디언트($\nabla_{W\_pred}L$)를 곱한 값($\eta \nabla_{W\_pred}L$)을 획득하고 학습률(에크, η)에 제2 그래디언트($\nabla_{W\_joint}L$)를 곱한 값($\eta \nabla_{W\_joint}L$)을 획득할 수 있다 (S1721).

[184] 또한, 전자 장치(100)는 제1 가중치 정보(W_pred-old) 및 값($\eta \nabla_{W\_pred}L$)에 기초하여 제1 서브 가중치 정보(W_pred-sub)를 획득하고, 제2 가중치 정보(W_joint-old) 및 값($\eta \nabla_{W\_joint}L$)에 기초하여 제2 서브 가중치 정보(W_joint-sub)를 획득할 수 있다 (S1722). 여기서, S1721, S1722 단계는 도 16의 S1620-1, S1620-2 단계에 대응될 수 있다.

[185] 또한, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub) 및 제2 서브 가중치 정보(W_joint-sub)의 평균값에 기초하여 업데이트된 제1 가중치 정보(W_pred-new)를 획득할 수 있다 (S1730-1).

[186] 또한, 전자 장치(100)는 제1 서브 가중치 정보(W_pred-sub) 및 제2 서브 가중치 정보(W_joint-sub)의 평균값에 기초하여 업데이트된 제2 가중치 정보(W_joint-new)를 획득할 수 있다 (S1730-2). 여기서, S1730-1, S1730-2 단계는 도 16의 S1630-1, S1630-2 단계에 대응될 수 있다.

[187] 도 18은 본 개시의 일 실시 예에 따른 전자 장치(100)의 제어 방법을 설명하기 위한 흐름도이다.

[188] 도 18을 참조하면, 복수의 네트워크로 구성된 음성 인식 모델 및 음성 인식 모델을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 저장하는 전자 장치(100)의 제어 방법은 제2 사용자 음성을 복수의 네트워크 중 제1 네트워크에 입력하여 제1 벡터를 획득하는 단계 (S1805), 제1 인식 정보를 복수의 네트워크

중 제1가중치 정보를 포함하는 제2 네트워크에 입력하여 제2 벡터를 획득하는 단계 (S1810) 및 제1 벡터 및 제2 벡터를 복수의 네트워크 중 제2 가중치 정보를 포함하는 제3 네트워크에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보를 획득하는 단계 (S1815)를 포함하고, 제2 가중치 정보 중 적어도 일부는 제1 가중치 정보와 동일한 정보일 수 있다.

[189] 한편, 음성 인식 모델은 RNN-T(Recurrent Neural Network Transducer) 모델일 수 있다.

[190] 한편, 제1 네트워크는 전사 네트워크(Transcription Network)이고, 제2 네트워크는 예측 네트워크(Prediction Network)이고, 제3 네트워크는 조인트 네트워크(Joint Network)일 수 있다.

[191] 한편, 제1 벡터를 획득하는 단계 (S1805)는 제2 사용자 음성이 수신되면, 제2 사용자 음성에 대응되는 특징 벡터를 획득하고, 제2 사용자 음성에 대응되는 특징 벡터 및 제1 네트워크에 포함된 제1 서브 네트워크에 기초하여 제1 벡터를 획득할 수 있다.

[192] 한편, 제2 벡터를 획득하는 단계 (S1810)는 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득하고, 제1 인식 정보에 대응되는 원-핫 벡터, 제1 가중치 정보 및 제2 네트워크에 포함된 제2 서브 네트워크에 기초하여 제2 벡터를 획득할 수 있다.

[193] 한편, 제2 인식 정보를 획득하는 단계 (S1815)는 제1 벡터, 제2 벡터 및 제3 네트워크에 포함된 제3 서브 네트워크에 기초하여 제3 벡터를 획득하고, 제3 벡터 및 제2 가중치 정보에 기초하여 제2 인식 정보를 획득할 수 있다.

[194] 한편, 제1 가중치 정보는 기 설정된 개수의 서브 워드에 대응되는 가중치를 포함하고, 제2 가중치 정보는 가중치 및 추가 가중치를 포함할 수 있다.

[195] 한편, 추가 가중치는 제2 사용자 음성에 대응되는 서브 워드가 존재하지 않는 경우 이용되는 가중치이고, 기 설정된 개수의 가중치의 차원(dimension)은 추가 가중치의 차원(dimension)과 동일할 수 있다.

[196] 한편, 제1 가중치 정보는 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트(gradient), 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트(gradient) 및 학습률(Learning rate)에 기초하여 학습되고, 제2 가중치 정보는 학습된 제1 가중치 정보에 기초하여 결정될 수 있다.

[197] 한편, 제1 가중치 정보 및 제2 가중치 정보는 제1 서브 가중치 정보 및 제2 서브 가중치 정보의 평균값에 기초하여 학습되고, 제1 서브 가중치는 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트 및 학습률에 기초하여 산출되고, 제2 서브 가중치는 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트 및 학습률에 기초하여 산출될 수 있다.

[198] 한편, 도 18과 같은 전자 장치의 제어 방법은 도 1의 구성을 가지는 전자 장치 상에서 실행될 수 있으며, 그 밖의 구성을 가지는 전자 장치 상에서도 실행될 수 있다.

[199] 한편, 상술한 본 개시의 다양한 실시 예들에 따른 방법들은, 기존 전자 장치에 설치 가능한 어플리케이션 형태로 구현될 수 있다.

[200] 또한, 상술한 본 개시의 다양한 실시 예들에 따른 방법들은, 기존 전자 장치에 대한 소프트웨어 업그레이드, 또는 하드웨어 업그레이드 만으로도 구현될 수 있다.

[201] 또한, 상술한 본 개시의 다양한 실시 예들은 전자 장치에 구비된 임베디드 서버, 또는 전자 장치 및 디스플레이 장치 중 적어도 하나의 외부 서버를 통해 수행되는 것도 가능하다.

[202] 한편, 본 개시의 일시 예에 따르면, 이상에서 설명된 다양한 실시 예들은 기기(machine)(예: 컴퓨터)로 읽을 수 있는 저장 매체(machine-readable storage media)에 저장된 명령어를 포함하는 소프트웨어로 구현될 수 있다. 기기는, 저장 매체로부터 저장된 명령어를 호출하고, 호출된 명령어에 따라 동작이 가능한 장치로서, 개시된 실시 예들에 따른 전자 장치를 포함할 수 있다. 명령이 프로세서에 의해 실행될 경우, 프로세서가 직접, 또는 프로세서의 제어 하에 다른 구성요소들을 이용하여 명령에 해당하는 기능을 수행할 수 있다. 명령은 컴파일러 또는 인터프리터에 의해 생성 또는 실행되는 코드를 포함할 수 있다. 기기로 읽을 수 있는 저장 매체는, 비일시적(non-transitory) 저장 매체의 형태로 제공될 수 있다. 여기서, '비일시적'은 저장 매체가 신호(signal)를 포함하지 않으며 실재(tangible)한다는 것을 의미할 뿐 데이터가 저장 매체에 반영구적 또는 임시적으로 저장됨을 구분하지 않는다.

[203] 또한, 본 개시의 일 실시 예에 따르면, 이상에서 설명된 다양한 실시 예들에 따른 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로, 또는 어플리케이션 스토어(예: 플레이 스토어TM)를 통해 온라인으로 배포될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

[204] 또한, 상술한 다양한 실시 예들에 따른 구성 요소(예: 모듈 또는 프로그램) 각각은 단수 또는 복수의 개체로 구성될 수 있으며, 전술한 해당 서브 구성 요소들 중 일부 서브 구성 요소가 생략되거나, 또는 다른 서브 구성 요소가 다양한 실시 예에 더 포함될 수 있다. 대체적으로 또는 추가적으로, 일부 구성 요소들(예: 모듈 또는 프로그램)은 하나의 개체로 통합되어, 통합되기 이전의 각각의 해당 구성 요소에 의해 수행되는 기능을 동일 또는 유사하게 수행할 수 있다. 다양한 실시 예들에 따른, 모듈, 프로그램 또는 다른 구성 요소에 의해 수행되는 동작들은 순차적, 병렬적, 반복적 또는 휴리스틱하게 실행되거나, 적어도 일부 동작이 다른 순서로 실행되거나, 생략되거나, 또는 다른 동작이

추가될 수 있다.

[205] 이상에서는 본 개시의 바람직한 실시 예에 대하여 도시하고 설명하였지만, 본 개시는 상술한 특정의 실시 예에 한정되지 아니하며, 청구범위에서 청구하는 본 개시의 요지를 벗어남이 없이 당해 개시에 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변형 실시가 가능한 것은 물론이고, 이러한 변형실시들은 본 개시의 기술적 사상이나 전망으로부터 개별적으로 이해되어져서는 안될 것이다.

청구범위

[청구항 1] 전자 장치에 있어서,
음성 인식 모델 및 상기 음성 인식 모델을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 저장하는 메모리, 상기 음성 인식 모델은 제1 네트워크, 제2 네트워크 및 제3 네트워크를 포함하고; 및
제2 사용자 음성에 대응되는 음성 데이터를 상기 제1 네트워크에 입력하여 제1 벡터를 획득하고,
상기 제1 인식 정보를 제1 가중치 정보에 기초하여 벡터를 생성하는 상기 제2 네트워크에 입력하여 제2 벡터를 획득하고,
상기 제1 벡터 및 상기 제2 벡터를 제2 가중치 정보에 기초하여 인식 정보를 생성하는 상기 제3 네트워크에 입력하여 상기 제2 사용자 음성에 대응되는 제2 인식 정보를 획득하는 프로세서;를 포함하고,
상기 제2 가중치 정보 중 적어도 일부는,
상기 제1 가중치 정보와 동일한, 전자 장치.

[청구항 2] 제1항에 있어서,
상기 음성 인식 모델은,
RNN-T(Recurrent Neural Network Transducer) 모델인, 전자 장치.

[청구항 3] 제2항에 있어서,
상기 제1 네트워크는 전사 네트워크(Transcription Network)이고,
상기 제2 네트워크는 예측 네트워크(Prediction Network)이고,
상기 제3 네트워크는 조인트 네트워크(Joint Network)인, 전자 장치.

[청구항 4] 제1항에 있어서,
상기 프로세서는,
상기 제2 사용자 음성이 수신되면, 상기 제2 사용자 음성에 대응되는 특징 벡터를 획득하고,
상기 제1 네트워크에 포함된 제1 서브 네트워크는 상기 특징 벡터에 기초하여 상기 제1 벡터를 생성하는, 전자 장치.

[청구항 5] 제1항에 있어서,
상기 프로세서는,
상기 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득하고,
상기 제2 네트워크에 포함된 제2 서브 네트워크는 상기 원-핫 벡터 및 상기 제1 가중치 정보에 기초하여 상기 제2 벡터를 생성하는, 전자 장치.

[청구항 6] 제1항에 있어서,
상기 프로세서는,
제3 네트워크에 포함된 제3 서브 네트워크는 상기 제1 벡터 및 상기 제2 벡터에 기초하여 제3 벡터를 획득하고,
상기 제3 네트워크는 상기 제3 벡터 및 상기 제2 가중치 정보에 기초하여

상기 제2 인식 정보를 생성하는, 전자 장치.

[청구항 7] 제1항에 있어서,
상기 제1 가중치 정보는 기 설정된 개수의 서브 워드에 대응되는 적어도 하나의 제1 가중치를 포함하고,
상기 제2 가중치 정보는 상기 적어도 하나의 제1 가중치 및 적어도 하나의 추가 가중치를 포함하고,
상기 적어도 하나의 제1 가중치는 상기 메모리의 제1 영역에 저장되고,
상기 적어도 하나의 추가 가중치는 상기 메모리의 제2 영역에 저장되고,
상기 프로세서는,
상기 제1 영역에 저장된 상기 적어도 하나의 제1 가중치 및 상기 제2 영역에 저장된 상기 적어도 하나의 가중치를 상기 제2 가중치 정보로써 이용하는, 전자 장치.

[청구항 8] 제7항에 있어서,
상기 적어도 하나의 추가 가중치는,
상기 제2 사용자 음성에 대응되는 상기 기 설정된 개수의 서브 워드가 존재하지 않는 경우 이용되는 가중치이고,
상기 적어도 하나의 제1 가중치의 차원(dimension)은 상기 적어도 하나의 추가 가중치의 차원(dimension)에 대응되는, 전자 장치.

[청구항 9] 제1항에 있어서,
상기 제1 가중치 정보는,
상기 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트(gradient), 상기 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트(gradient) 및 학습률(Learning rate)에 기초하여 학습되고,
상기 제2 가중치 정보는,
상기 학습된 제1 가중치 정보에 기초하여 결정되는, 전자 장치.

[청구항 10] 제1항에 있어서,
상기 제1 가중치 정보 및 상기 제2 가중치 정보 각각은,
제1 서브 가중치 정보 및 제2 서브 가중치 정보의 평균값에 기초하여 학습되고,
상기 제1 서브 가중치 정보는,
상기 제1 가중치 정보에 따른 손실값의 변화량을 나타내는 제1 그래디언트 및 학습률에 기초하여 결정되고,
상기 제2 서브 가중치 정보는,
상기 제2 가중치 정보에 따른 손실값의 변화량을 나타내는 제2 그래디언트 및 상기 학습률에 기초하여 결정되는, 전자 장치.

[청구항 11] 음성 인식 모델 및 상기 음성 인식 모델을 통해 획득한 제1 사용자 음성에 대응되는 제1 인식 정보를 저장하는 전자 장치의 제어 방법에 있어서,

상기 음성 인식 모델은 제1 네트워크, 제2 네트워크 및 제3 네트워크를 포함하고,
제2 사용자 음성에 대응되는 음성 데이터를 상기 제1 네트워크에 입력하여 제1 벡터를 획득하는 단계;
상기 제1 인식 정보를 제1 가중치 정보에 기초하여 벡터를 생성하는 상기 제2 네트워크에 입력하여 제2 벡터를 획득하는 단계; 및
상기 제1 벡터 및 상기 제2 벡터를 제2 가중치 정보에 기초하여 인식 정보를 생성하는 상기 제3 네트워크에 입력하여 상기 제2 사용자 음성에 대응되는 제2 인식 정보를 획득하는 단계;를 포함하고,
상기 제2 가중치 정보 중 적어도 일부는,
상기 제1 가중치 정보와 동일한, 제어 방법.

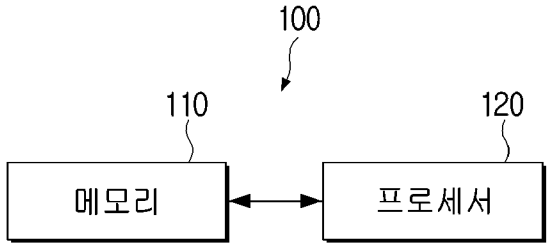
[청구항 12] 제11항에 있어서,
상기 음성 인식 모델은,
RNN-T(Recurrent Neural Network Transducer) 모델인, 제어 방법.

[청구항 13] 제12항에 있어서,
상기 제1 네트워크는 전사 네트워크(Transcription Network)이고,
상기 제2 네트워크는 예측 네트워크(Prediction Network)이고,
상기 제3 네트워크는 조인트 네트워크(Joint Network)인, 제어 방법.

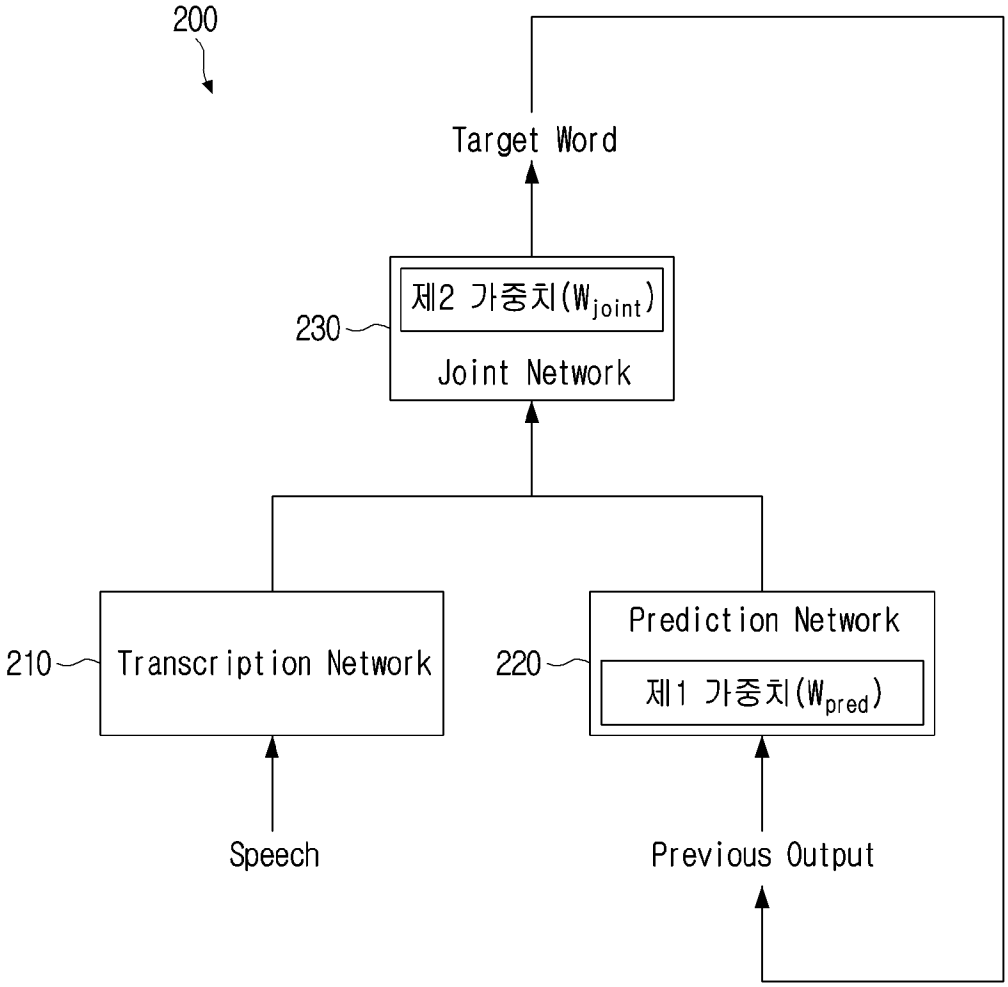
[청구항 14] 제11항에 있어서,
상기 제1 벡터를 획득하는 단계는,
상기 제2 사용자 음성이 수신되면, 상기 제2 사용자 음성에 대응되는 특징 벡터를 획득하고,
상기 제1 네트워크에 포함된 제1 서브 네트워크는 상기 특징 벡터에 기초하여 상기 제1 벡터를 생성하는, 제어 방법.

[청구항 15] 제11항에 있어서,
상기 제2 벡터를 획득하는 단계는,
상기 제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득하고,
상기 제2 네트워크에 포함된 제2 서브 네트워크는 상기 원-핫 벡터 및 상기 제1 가중치 정보에 기초하여 상기 제2 벡터를 생성하는, 제어 방법.

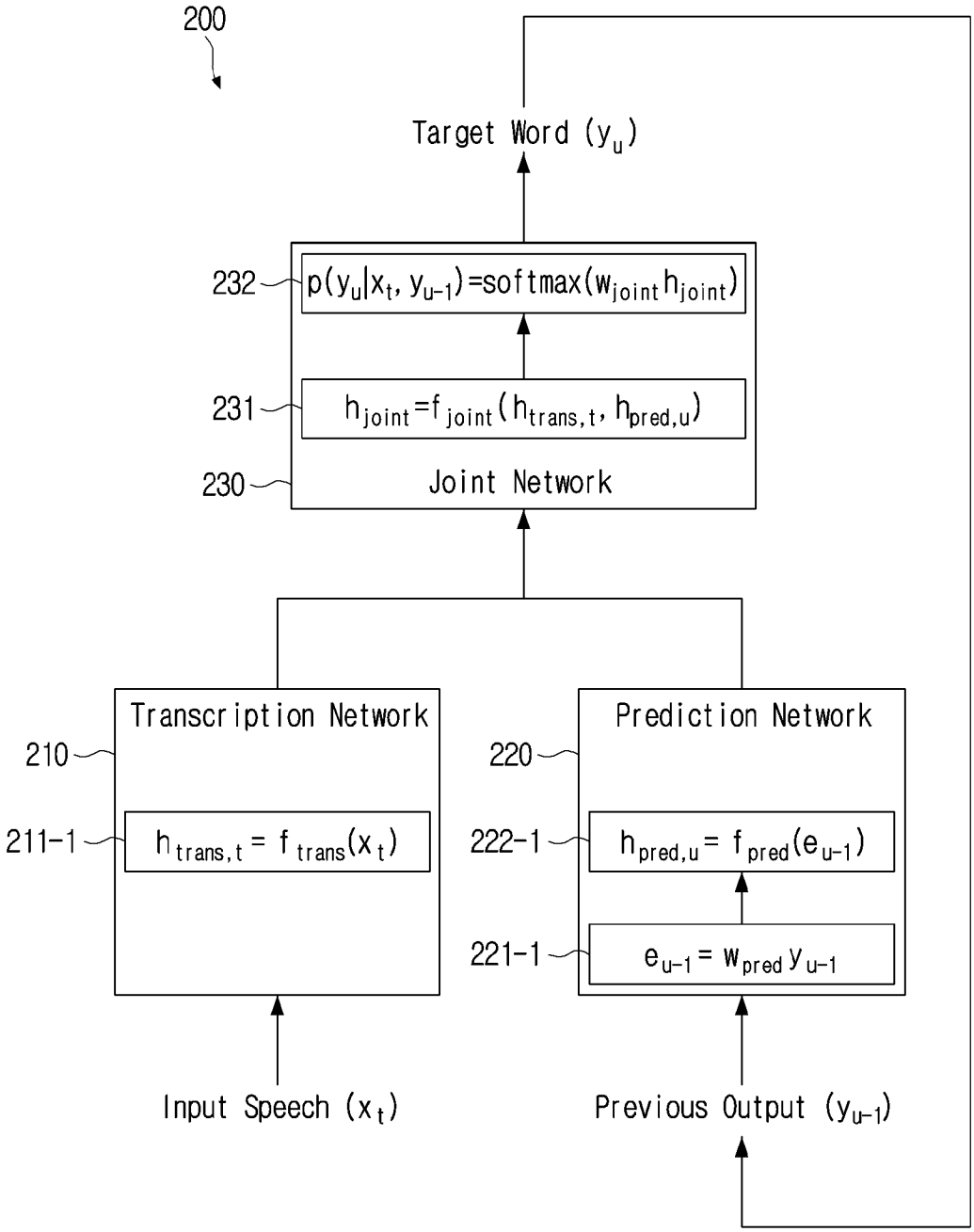
[도1]


100
110
120
메모리
프로세서

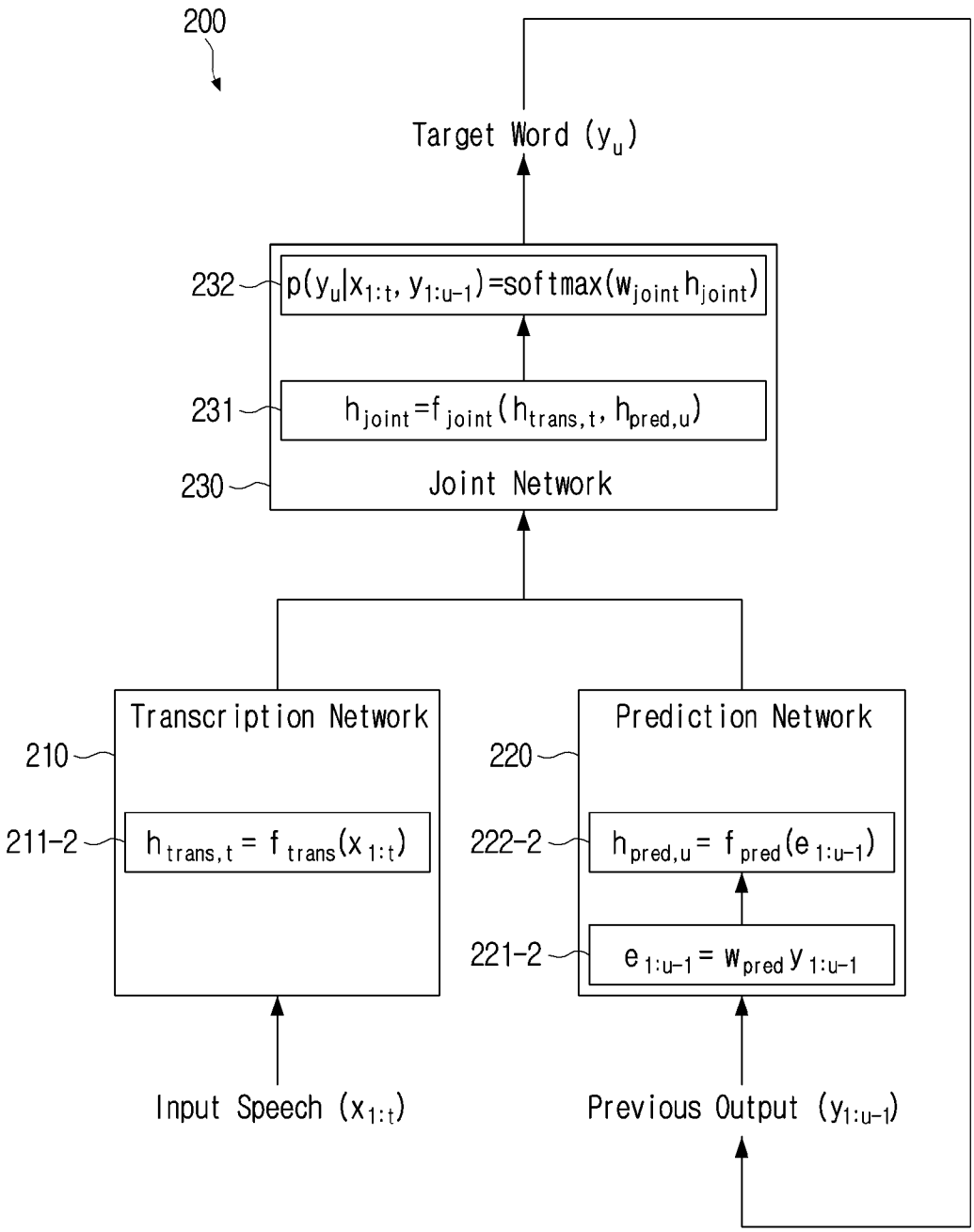
[도2]


200
Target Word
제2 가중치(W_{joint})
230
Joint Network
Transcription Network
210
Prediction Network
220
제1 가중치(W_{pred})
Speech
Previous Output

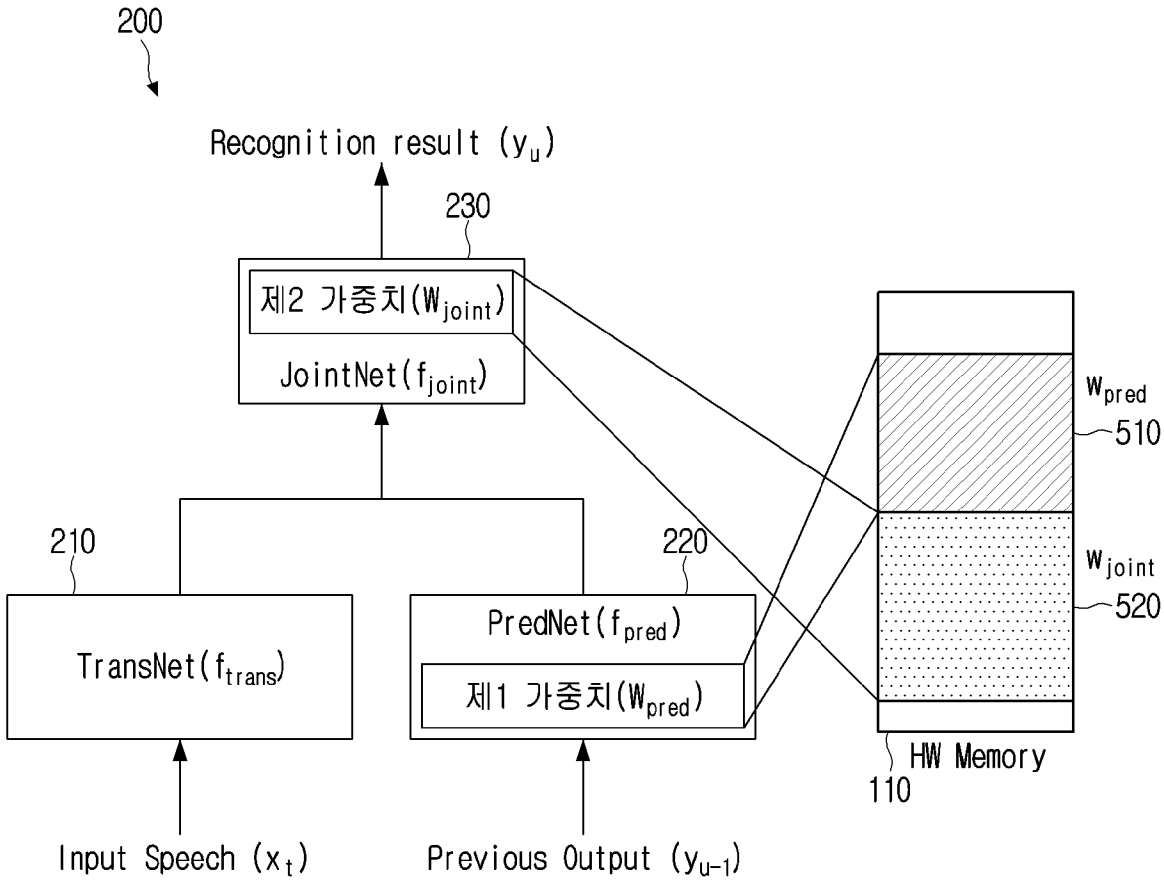
[도3]


200
Target Word (y_u)
232
$p(y_u|x_t, y_{u-1})=softmax(w_{joint} h_{joint})$
231
$h_{joint}=f_{joint}(h_{trans,t}, h_{pred,u})$
230
Joint Network
Transcription Network
210
211-1
$h_{trans,t} = f_{trans}(x_t)$
Prediction Network
220
222-1
$h_{pred,u} = f_{pred}(e_{u-1})$
221-1
$e_{u-1} = w_{pred} y_{u-1}$
Input Speech (x_t)
Previous Output (y_{u-1})

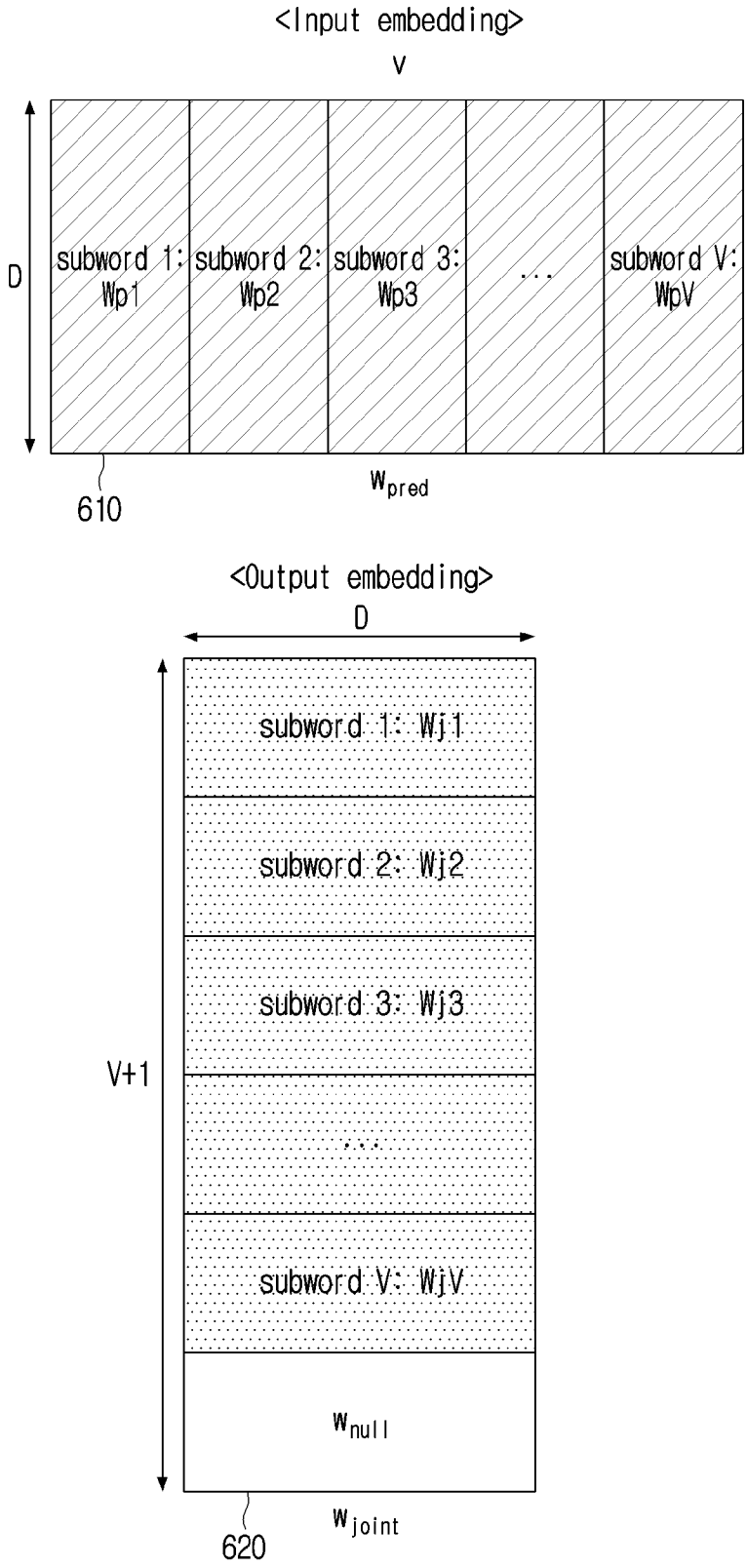
[도4]


200
Target Word (y_u)
232
$p(y_u|x_{1:t}, y_{1:u-1})=softmax(w_{joint}h_{joint})$
231
$h_{joint}=f_{joint}(h_{trans,t}, h_{pred,u})$
230
Joint Network
Transcription Network
210
211-2
$h_{trans,t} = f_{trans}(x_{1:t})$
Prediction Network
220
222-2
$h_{pred,u} = f_{pred}(e_{1:u-1})$
221-2
$e_{1:u-1} = w_{pred} y_{1:u-1}$
Input Speech ($x_{1:t}$)
Previous Output ($y_{1:u-1}$)

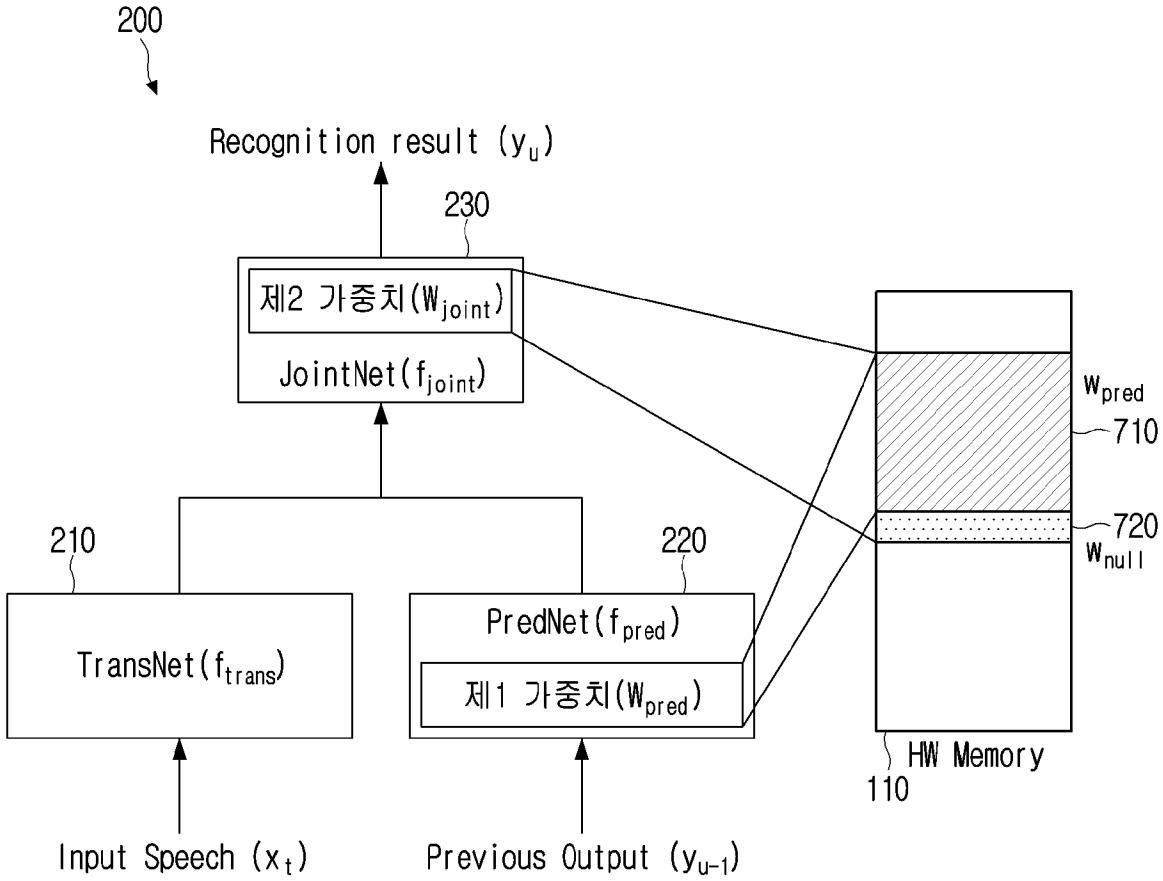
[도5]


200
Recognition result (y_u)
230
제2 가중치(W_{joint})
JointNet(f_{joint})
w_{pred}
510
w_{joint}
520
210
220
TransNet(f_{trans})
PredNet(f_{pred})
제1 가중치(W_{pred})
HW Memory
110
Input Speech (x_t)
Previous Output (y_{u-1})

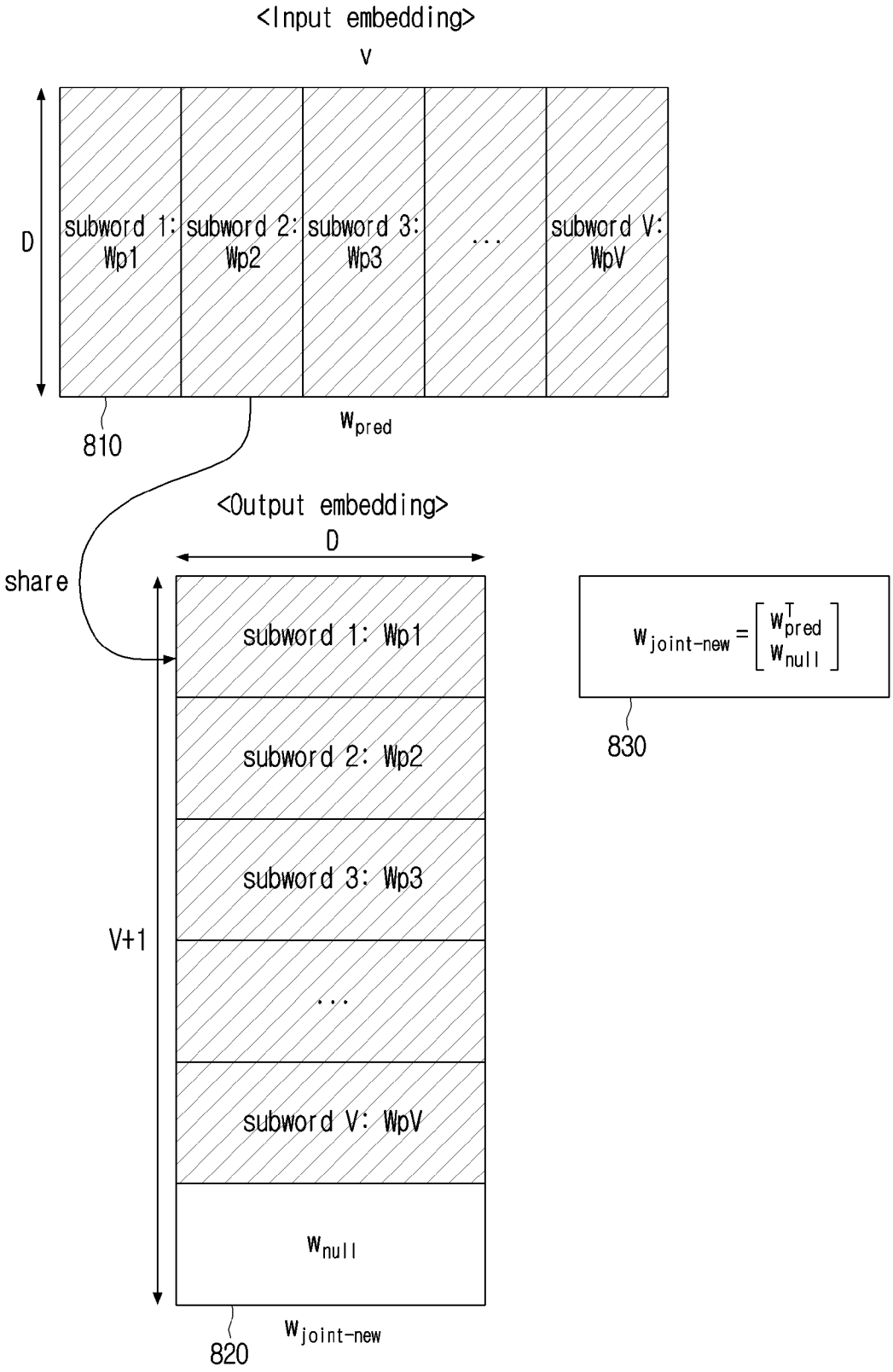
[도6]


<Input embedding>
V
D
subword 1: Wp1
subword 2: Wp2
subword 3: Wp3
...
subword V: WpV
610
W_{pred}
<Output embedding>
D
subword 1: Wj1
subword 2: Wj2
subword 3: Wj3
V+1
...
subword V: WjV
W_{null}
620
W_{joint}

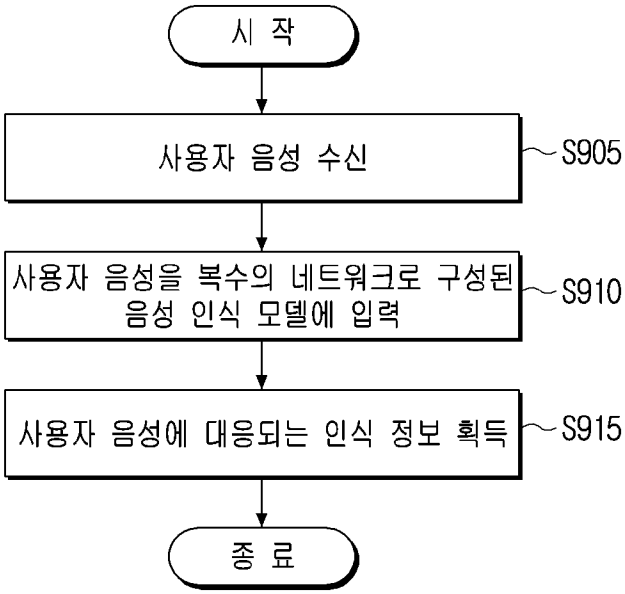
[도7]


200
Recognition result (y_u)
230
제2 가중치(W_{joint})
JointNet(f_{joint})
W_{pred}
710
720
W_{null}
210
220
TransNet(f_{trans})
PredNet(f_{pred})
제1 가중치(W_{pred})
HW Memory
110
Input Speech (x_t)
Previous Output (y_{u-1})

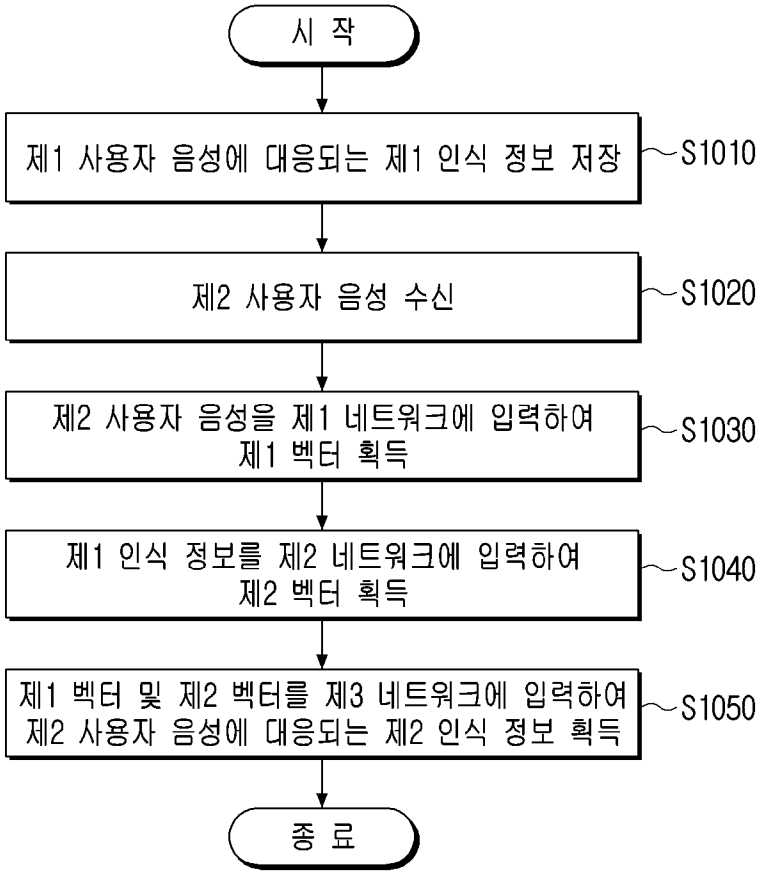
[도8]


<Input embedding>
V
D
subword 1: Wp1
subword 2: Wp2
subword 3: Wp3
...
subword V: WpV
810
w_{pred}
<Output embedding>
D
share
subword 1: Wp1
subword 2: Wp2
subword 3: Wp3
...
subword V: WpV
w_{null}
V+1
820
$w_{joint-new}$
$w_{joint-new} = \begin{bmatrix} w_{pred}^T \\ w_{null} \end{bmatrix}$
830

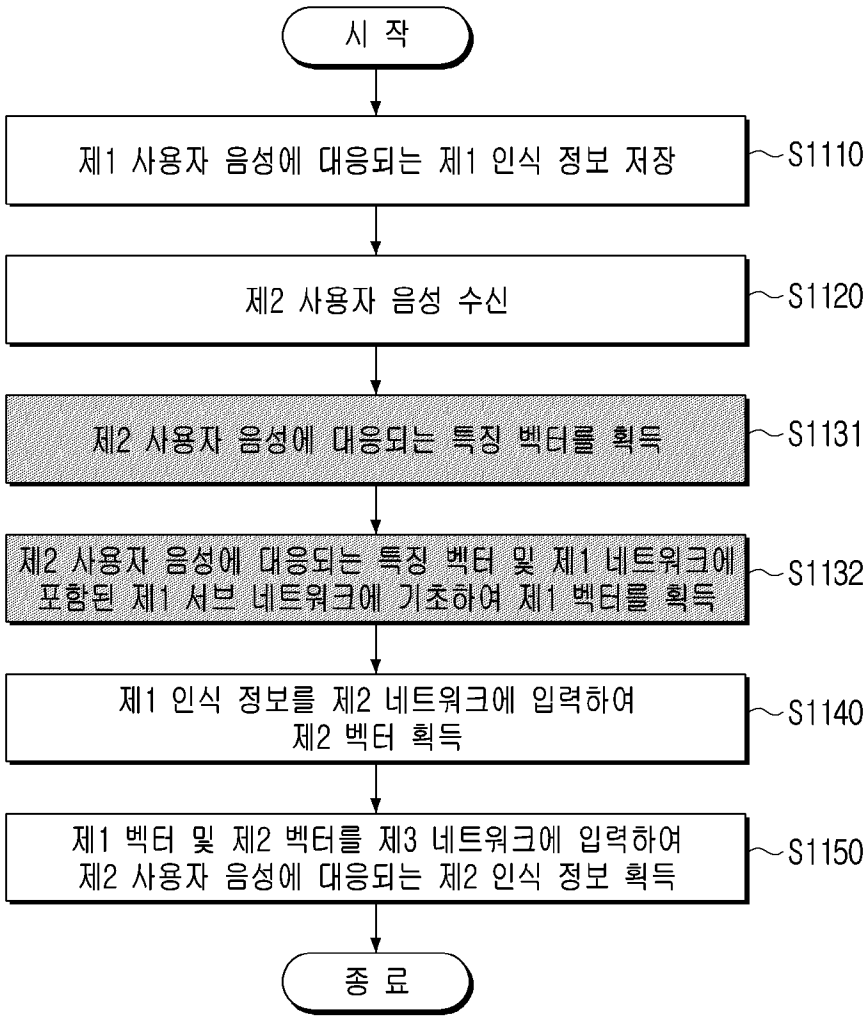
[도9]


시 작
사용자 음성 수신
S905
사용자 음성을 복수의 네트워크로 구성된 음성 인식 모델에 입력
S910
사용자 음성에 대응되는 인식 정보 획득
S915
종 료

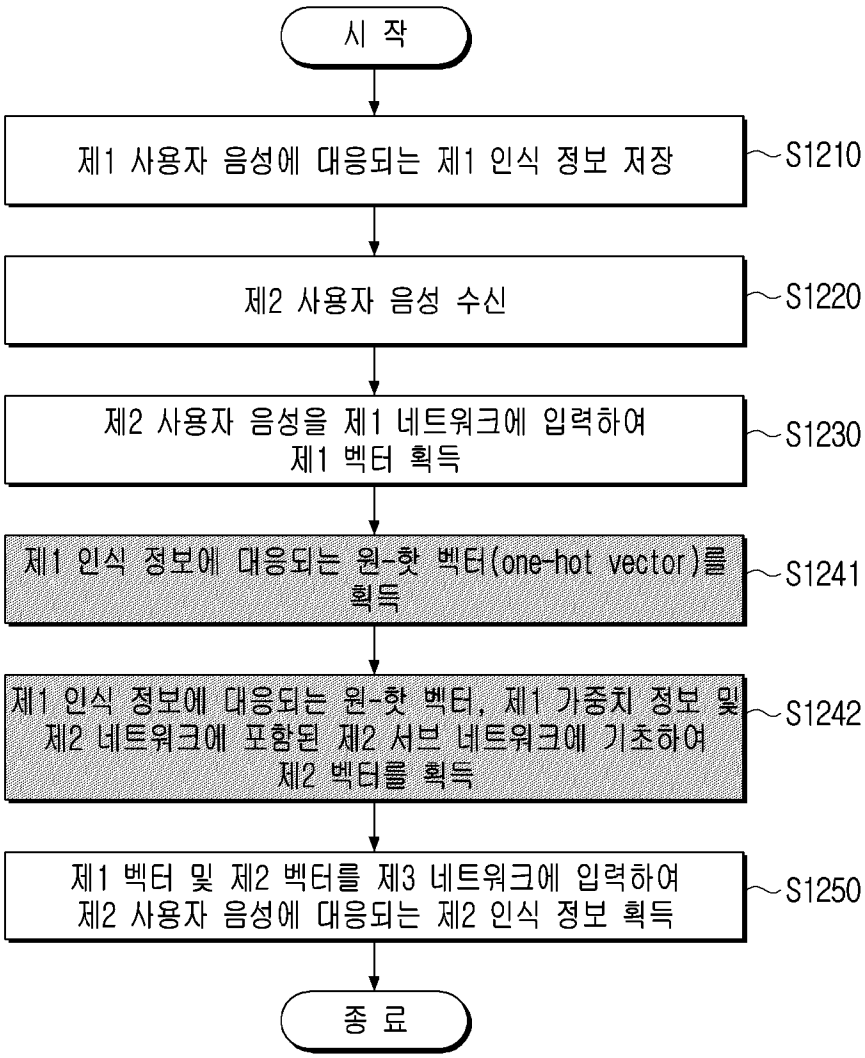
[도10]


시 작
제1 사용자 음성에 대응되는 제1 인식 정보 저장
S1010
제2 사용자 음성 수신
S1020
제2 사용자 음성을 제1 네트워크에 입력하여 제1 벡터 획득
S1030
제1 인식 정보를 제2 네트워크에 입력하여 제2 벡터 획득
S1040
제1 벡터 및 제2 벡터를 제3 네트워크에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보 획득
S1050
종 료

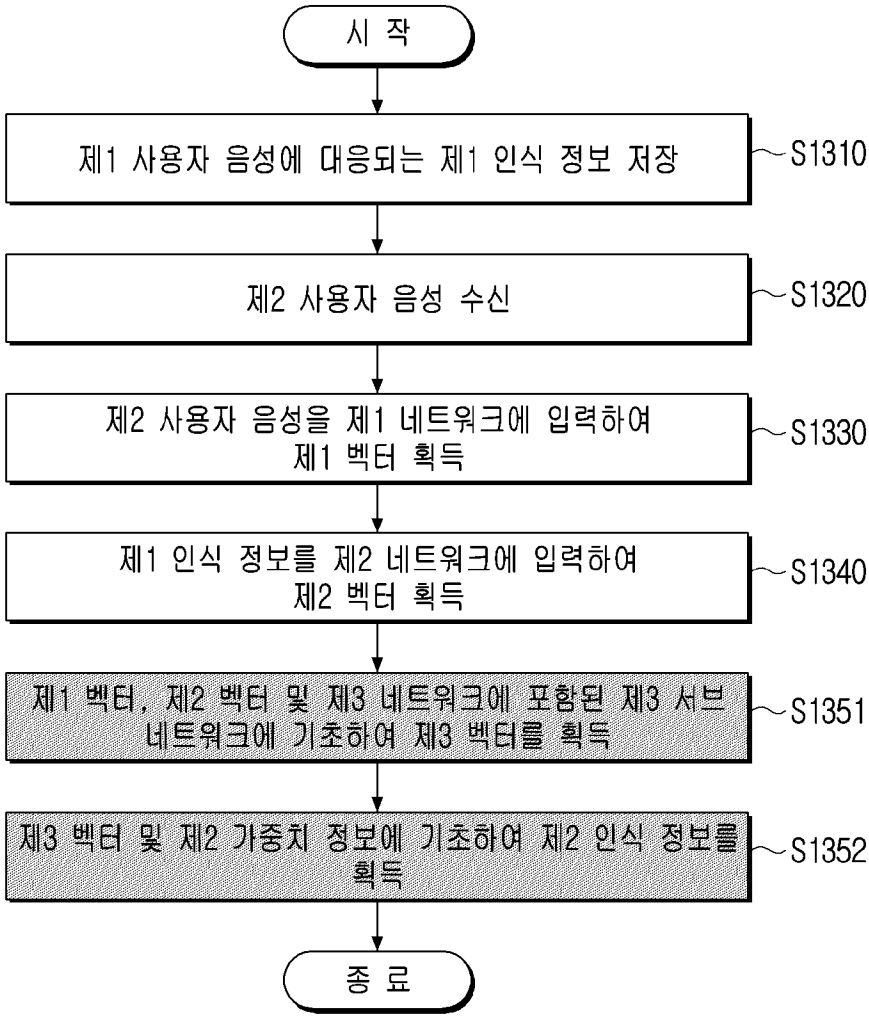
[도11]


시 작
제1 사용자 음성에 대응되는 제1 인식 정보 저장
S1110
제2 사용자 음성 수신
S1120
제2 사용자 음성에 대응되는 특징 벡터를 획득
S1131
제2 사용자 음성에 대응되는 특징 벡터 및 제1 네트워크에 포함된 제1 서브 네트워크에 기초하여 제1 벡터를 획득
S1132
제1 인식 정보를 제2 네트워크에 입력하여 제2 벡터 획득
S1140
제1 벡터 및 제2 벡터를 제3 네트워크에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보 획득
S1150
종 료

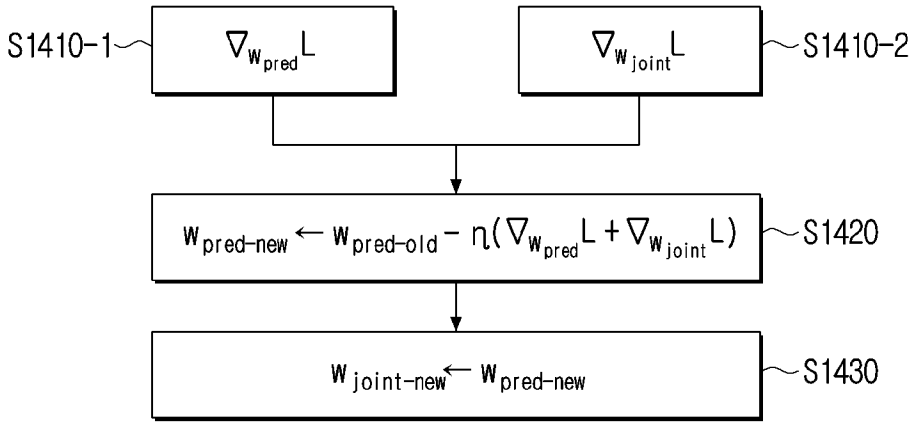
[도12]


시 작
제1 사용자 음성에 대응되는 제1 인식 정보 저장
S1210
제2 사용자 음성 수신
S1220
제2 사용자 음성을 제1 네트워크에 입력하여 제1 벡터 획득
S1230
제1 인식 정보에 대응되는 원-핫 벡터(one-hot vector)를 획득
S1241
제1 인식 정보에 대응되는 원-핫 벡터, 제1 가중치 정보 및 제2 네트워크에 포함된 제2 서브 네트워크에 기초하여 제2 벡터를 획득
S1242
제1 벡터 및 제2 벡터를 제3 네트워크에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보 획득
S1250
종 료

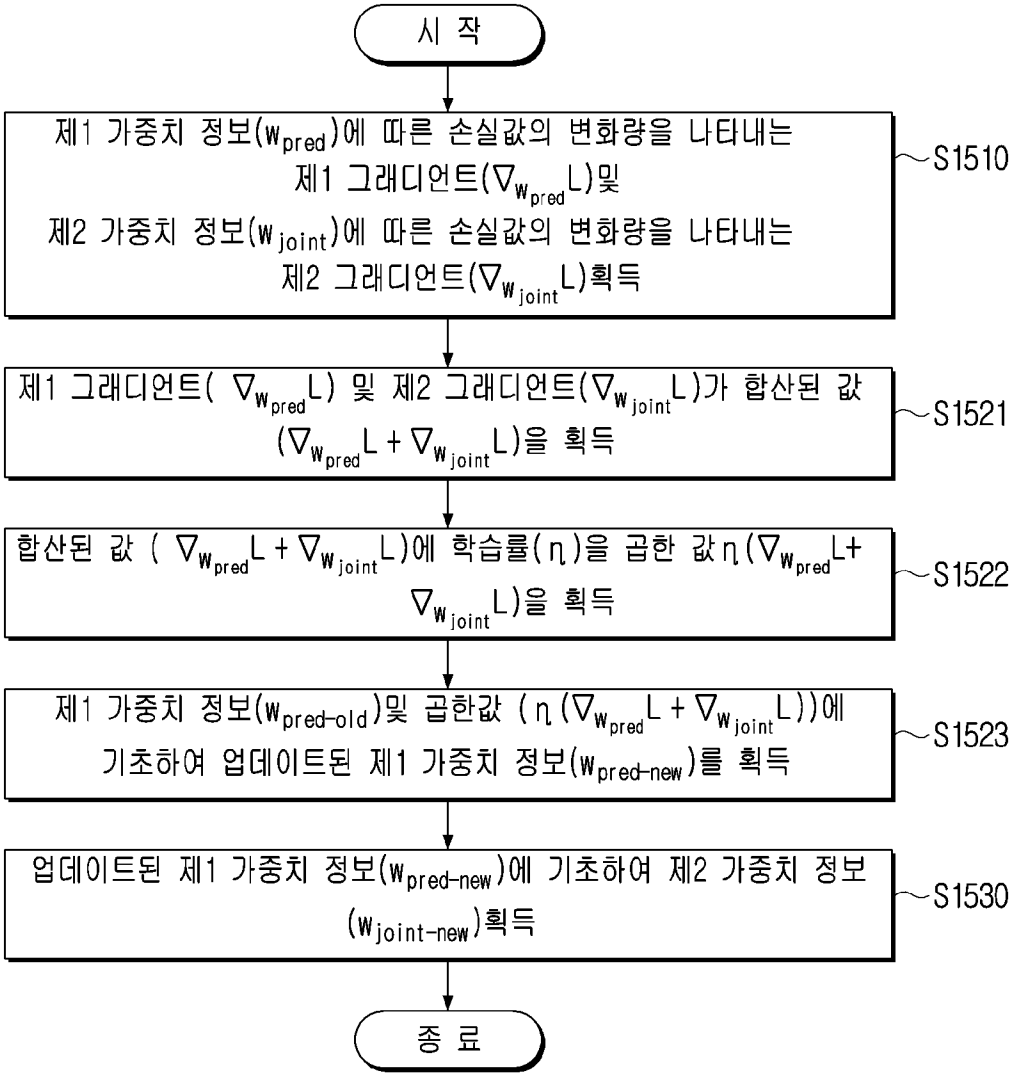
[도13]


시 작
제1 사용자 음성에 대응되는 제1 인식 정보 저장
S1310
제2 사용자 음성 수신
S1320
제2 사용자 음성을 제1 네트워크에 입력하여 제1 벡터 획득
S1330
제1 인식 정보를 제2 네트워크에 입력하여 제2 벡터 획득
S1340
제1 벡터, 제2 벡터 및 제3 네트워크에 포함된 제3 서브 네트워크에 기초하여 제3 벡터를 획득
S1351
제3 벡터 및 제2 가중치 정보에 기초하여 제2 인식 정보를 획득
S1352
종 료

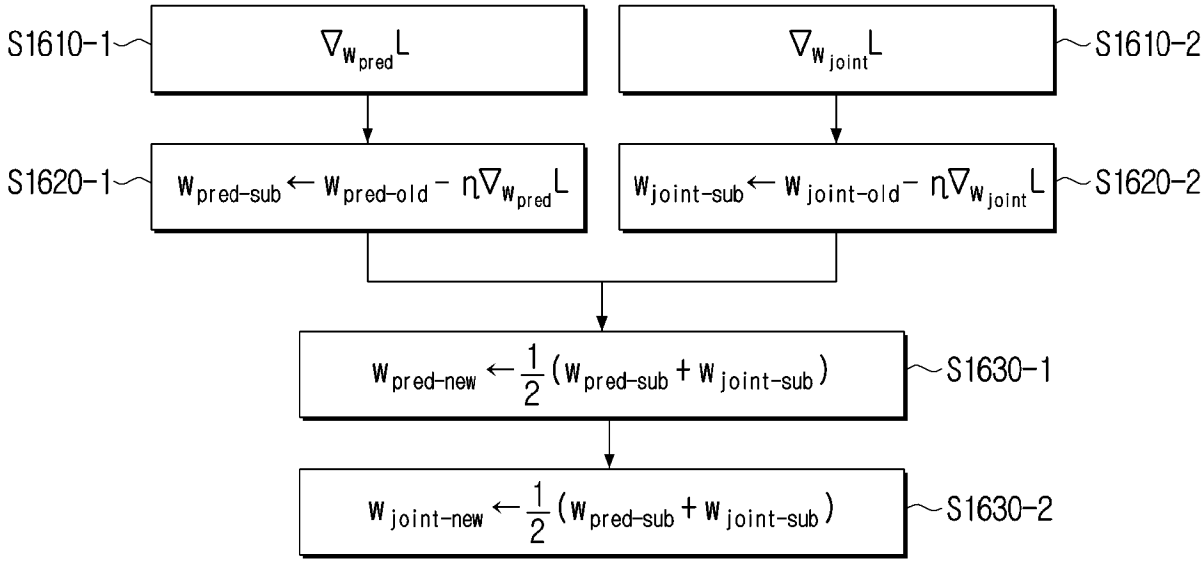
[도14]


S1410-1
$\nabla_{W_{pred}} L$
$\nabla_{W_{joint}} L$
S1410-2
$W_{pred\text{-}new} \leftarrow W_{pred\text{-}old} - \eta(\nabla_{W_{pred}} L + \nabla_{W_{joint}} L)$
S1420
$W_{joint\text{-}new} \leftarrow W_{pred\text{-}new}$
S1430

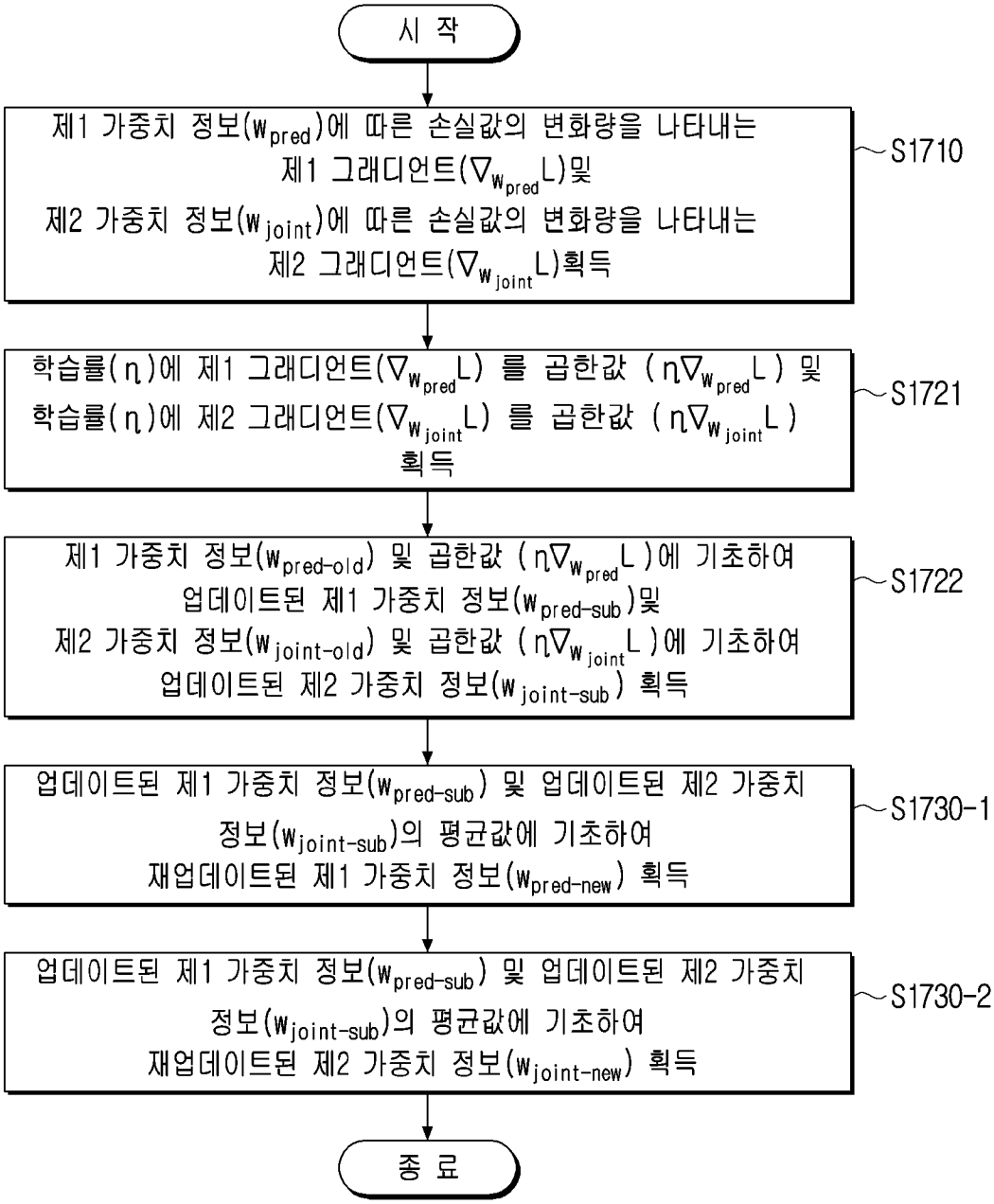
[도15]


시 작
제1 가중치 정보(w_{pred})에 따른 손실값의 변화량을 나타내는 제1 그래디언트($\nabla_{w_{pred}}L$)및 제2 가중치 정보(w_{joint})에 따른 손실값의 변화량을 나타내는 제2 그래디언트($\nabla_{w_{joint}}L$)획득
S1510
제1 그래디언트($\nabla_{w_{pred}}L$) 및 제2 그래디언트($\nabla_{w_{joint}}L$)가 합산된 값 ($\nabla_{w_{pred}}L+\nabla_{w_{joint}}L$)을 획득
S1521
합산된 값 ($\nabla_{w_{pred}}L+\nabla_{w_{joint}}L$)에 학습률($\eta$)을 곱한 값$\eta(\nabla_{w_{pred}}L+\nabla_{w_{joint}}L$)을 획득
S1522
제1 가중치 정보($w_{pred-old}$)및 곱한값 ($\eta(\nabla_{w_{pred}}L+\nabla_{w_{joint}}L)$)에 기초하여 업데이트된 제1 가중치 정보($w_{pred-new}$)를 획득
S1523
업데이트된 제1 가중치 정보($w_{pred-new}$)에 기초하여 제2 가중치 정보 ($w_{joint-new}$)획득
S1530
종 료

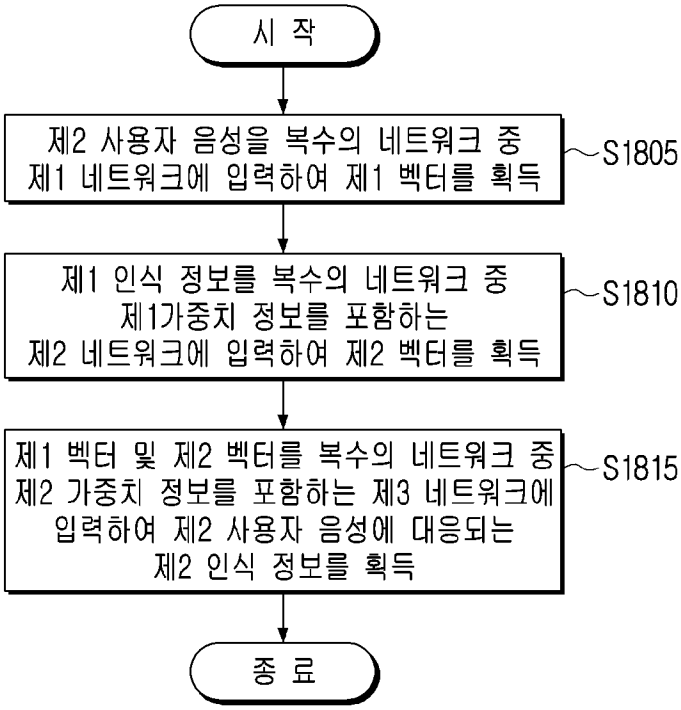
[도16]


S1610-1
$\nabla_{w_{pred}}L$
$\nabla_{w_{joint}}L$
S1610-2
S1620-1
$w_{pred-sub} \leftarrow w_{pred-old} - \eta\nabla_{w_{pred}}L$
$w_{joint-sub} \leftarrow w_{joint-old} - \eta\nabla_{w_{joint}}L$
S1620-2
$w_{pred-new} \leftarrow \frac{1}{2}(w_{pred-sub} + w_{joint-sub})$
S1630-1
$w_{joint-new} \leftarrow \frac{1}{2}(w_{pred-sub} + w_{joint-sub})$
S1630-2

[도17]


시 작
제1 가중치 정보(w_{pred})에 따른 손실값의 변화량을 나타내는 제1 그래디언트($\nabla_{w_{pred}}L$)및 제2 가중치 정보(w_{joint})에 따른 손실값의 변화량을 나타내는 제2 그래디언트($\nabla_{w_{joint}}L$)획득
S1710
학습률(η)에 제1 그래디언트($\nabla_{w_{pred}}L$) 를 곱한값 ($\eta\nabla_{w_{pred}}L$) 및 학습률(η)에 제2 그래디언트($\nabla_{w_{joint}}L$) 를 곱한값 ($\eta\nabla_{w_{joint}}L$) 획득
S1721
제1 가중치 정보($w_{pred-old}$) 및 곱한값 ($\eta\nabla_{w_{pred}}L$)에 기초하여 업데이트된 제1 가중치 정보($w_{pred-sub}$)및 제2 가중치 정보($w_{joint-old}$) 및 곱한값 ($\eta\nabla_{w_{joint}}L$)에 기초하여 업데이트된 제2 가중치 정보($w_{joint-sub}$) 획득
S1722
업데이트된 제1 가중치 정보($w_{pred-sub}$) 및 업데이트된 제2 가중치 정보($w_{joint-sub}$)의 평균값에 기초하여 재업데이트된 제1 가중치 정보($w_{pred-new}$) 획득
S1730-1
업데이트된 제1 가중치 정보($w_{pred-sub}$) 및 업데이트된 제2 가중치 정보($w_{joint-sub}$)의 평균값에 기초하여 재업데이트된 제2 가중치 정보($w_{joint-new}$) 획득
S1730-2
종 료

[도18]


시 작
제2 사용자 음성을 복수의 네트워크 중 제1 네트워크에 입력하여 제1 벡터를 획득
S1805
제1 인식 정보를 복수의 네트워크 중 제1가중치 정보를 포함하는 제2 네트워크에 입력하여 제2 벡터를 획득
S1810
제1 벡터 및 제2 벡터를 복수의 네트워크 중 제2 가중치 정보를 포함하는 제3 네트워크에 입력하여 제2 사용자 음성에 대응되는 제2 인식 정보를 획득
S1815
종 료

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2022/013533

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/12(2006.01)i; **G10L 15/06**(2006.01)i; **G10L 15/28**(2006.01)i; **G10L 15/22**(2006.01)i; **G06F 3/06**(2006.01)i; **G06F 3/16**(2006.01)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 15/12(2006.01); G10L 15/02(2006.01); G10L 15/06(2006.01); G10L 15/10(2006.01); G10L 15/16(2006.01); G10L 15/183(2013.01); G10L 17/04(2013.01); G10L 19/035(2013.01); G10L 25/30(2013.01)

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) & keywords: 음성인식모델(voice recognition model), 네트워크(network), 가중치(weight value), 벡터(vector), RNN-T(Recurrent Neural Network Transducer)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2018-0061397 A1 (ALIBABA GROUP HOLDING LIMITED) 01 March 2018 (2018-03-01) See paragraphs [0019] and [0125]-[0148]; claims 1-9; and figures 7-9.	1-15
A	US 9984682 B1 (EDUCATIONAL TESTING SERVICE) 29 May 2018 (2018-05-29) See column 14, line 16 - column 15, line 52; claims 1-6; and figures 8-12.	1-15
A	KR 10-2018-0018031 A (ELECTRONICS AND TELECOMMUNICATIONS RESEARCH INSTITUTE) 21 February 2018 (2018-02-21) See paragraphs [0029]-[0049]; claims 1-12; and figures 1-2.	1-15
A	KR 10-2017-0119152 A (IUCF-HYU (INDUSTRY-UNIVERSITY COOPERATION FOUNDATION HANYANG UNIVERSITY)) 26 October 2017 (2017-10-26) See paragraphs [0052]-[0090]; claims 1-7; and figure 2.	1-15
A	CN 110675859 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 10 January 2020 (2020-01-10) See paragraphs [0049]-[0092]; claims 1-7; and figures 1 and 3.	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
"D" document cited by the applicant in the international application
"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
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 December 2022	**22 December 2022**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **Government Complex-Daejeon Building 4, 189 Cheongsa-ro, Seo-gu, Daejeon 35208**	
Facsimile No. **+82-42-481-8578**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2022/013533

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2018-0061397	A1	01 March 2018	CN	107785015	A	09 March 2018
				EP	3504703	A1	03 July 2019
				EP	3504703	B1	03 August 2022
				JP	2019-528476	A	10 October 2019
				JP	7023934	B2	22 February 2022
				WO	2018-039500	A1	01 March 2018
US	9984682	B1	29 May 2018	US	10559225	B1	11 February 2020
KR	10-2018-0018031	A	21 February 2018	KR	10-2033411	B1	17 October 2019
				US	2018-0047389	A1	15 February 2018
KR	10-2017-0119152	A	26 October 2017	None			
CN	110675859	A	10 January 2020	CN	110675859	B	23 November 2021

국 제 조 사 보 고 서

국제출원번호
PCT/KR2022/013533

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G10L 15/12(2006.01)i; **G10L 15/06**(2006.01)i; **G10L 15/28**(2006.01)i; **G10L 15/22**(2006.01)i; **G06F 3/06**(2006.01)i; **G06F 3/16**(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G10L 15/12(2006.01); G10L 15/02(2006.01); G10L 15/06(2006.01); G10L 15/10(2006.01); G10L 15/16(2006.01); G10L 15/183(2013.01); G10L 17/04(2013.01); G10L 19/035(2013.01); G10L 25/30(2013.01)

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: 음성인식모델(voice recognition model), 네트워크(network), 가중치(weight value), 벡터(vector), RNN-T(Recurrent Neural Network Transducer)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	US 2018-0061397 A1 (ALIBABA GROUP HOLDING LIMITED) 2018.03.01 단락 [0019], [0125]-[0148]; 청구항 1-9; 및 도면 7-9	1-15
A	US 9984682 B1 (EDUCATIONAL TESTING SERVICE) 2018.05.29 컬럼 14, 라인 16 - 컬럼 15, 라인 52; 청구항 1-6; 및 도면 8-12	1-15
A	KR 10-2018-0018031 A (한국전자통신연구원) 2018.02.21 단락 [0029]-[0049]; 청구항 1-12; 및 도면 1-2	1-15
A	KR 10-2017-0119152 A (한양대학교 산학협력단) 2017.10.26 단락 [0052]-[0090]; 청구항 1-7; 및 도면 2	1-15
A	CN 110675859 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 2020.01.10 단락 [0049]-[0092]; 청구항 1-7; 및 도면 1, 3	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 이후에 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	국제조사보고서 발송일
2022년12월22일(22.12.2022)	**2022년12월22일(22.12.2022)**

ISA/KR의 명칭 및 우편주소	심사관
대한민국 특허청 **(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)**	**고재용**
팩스 번호 **+82-42-481-8578**	전화번호 **+82-42-481-8212**

국 제 조 사 보 고 서
대응특허에 관한 정보

국제출원번호
PCT/KR2022/013533

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
US 2018-0061397 A1	2018/03/01	CN 107785015 A	2018/03/09
		EP 3504703 A1	2019/07/03
		EP 3504703 B1	2022/08/03
		JP 2019-528476 A	2019/10/10
		JP 7023934 B2	2022/02/22
		WO 2018-039500 A1	2018/03/01
US 9984682 B1	2018/05/29	US 10559225 B1	2020/02/11
KR 10-2018-0018031 A	2018/02/21	KR 10-2033411 B1	2019/10/17
		US 2018-0047389 A1	2018/02/15
KR 10-2017-0119152 A	2017/10/26	없음	
CN 110675859 A	2020/01/10	CN 110675859 B	2021/11/23