

(51) 국제특허분류:

G10L 13/08 (2006.01)

G10L 25/18 (2013.01)

G10L 13/02 (2006.01)

(21) 국제출원번호:

PCT/KR2021/095116

(22) 국제출원일:

2021 년 12 월 2 일 (02.12.2021)

(25) 출원언어:

한국어

(26) 공개언어:

한국어

(30) 우선권정보:

10-2021-0004856 2021 년 1 월 13 일 (13.01.2021) KR

(71) 출원인:

한양대학교산학협력단

(IN-

 DUSTY-UNIVERSITY COOPERATION FOUNDATION HANYANG UNIVERSITY) [KR/KR]; 04763
 서울시 성동구 왕십리로 222, Seoul (KR).

 (72) 발명자: 장준혁 (CHANG, Joon Hyuk); 04763 서울시
 성동구 왕십리로 222, Seoul (KR). 황성웅 (HWANG,
 Sung Woong); 04763 서울시 성동구 왕십리로 222, Seoul
 (KR).

 (74) 대리인: 심경식 등 (SHIM, Kyoung-Shik et al.); 06227 서
 울시 강남구 역삼로 212, 6층, Seoul (KR).

 (81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국
 내 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, IT, 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.

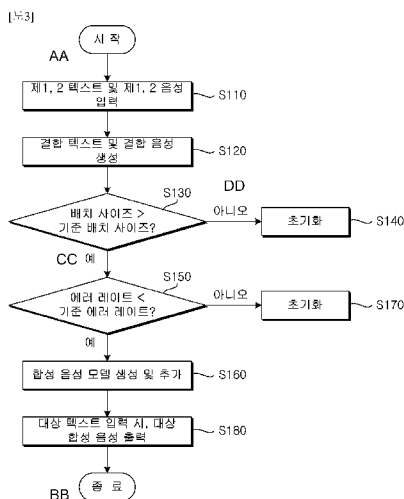
 (84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역
 내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE,
 LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM,
 ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유
 럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
 FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK,
 MT, NL, NO, PL, PT,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

(54) Title: OPERATION METHOD OF SPEECH SYNTHESIS SYSTEM

(54) 발명의 명칭: 음성 합성 시스템의 동작방법



S110 ... Input first and second texts, and first and second speeches
 S120 ... Generate combined text and combined speech
 S130 ... Arrangement size > Reference arrangement size?
 S140, S170 ... Initialize
 S150 ... Error rate < Reference error rate?
 S160 ... Generate speech synthesis model and add
 S180 ... Output target synthetic speech when target text is input
 AA ... Start
 BB ... End
 CC ... Yes
 DD ... No

(57) Abstract: The present invention provides an operation method of a speech synthesis system, comprising the steps of: inputting a first text, a first speech for the first text, a second text, and a second speech for the second text; generating a speech synthesis model trained by applying the first and second texts and the first and second speeches to curriculum learning; and outputting a target synthetic speech corresponding to a target text, on the basis of the speech synthesis model, when the target text for a speech output is input, wherein the step for generating the speech synthesis model comprises the steps of: generating a combined text in which the first and second texts are combined and a combined speech in which the first and second speeches are combined; and adding the combined text and the combined speech to the speech synthesis model when an error rate during learning combination of the combined text and combined speech is smaller than a reference rate.

(57) 요약서: 본 발명은, 제1 텍스트와 상기 제1 텍스트에 대한 제1 음성 및 제2 텍스트와 제2 텍스트에 대한 제2 음성이 입력되는 단계, 제1, 2 텍스트 및 제1, 2 음성을 커리큘럼 러닝에 적용하여 학습한 음성 합성 모델을 생성하는 단계 및 음성 출력을 위한 대상 텍스트 입력 시, 음성 합성 모델을 기반으로 대상 텍스트에 대응하는 대상 합성 음성을 출력하는 단계를 포함하고, 음성 합성 모델을 생성하는 단계는, 제1, 2 텍스트를 결합한 결합 텍스트 및 제1, 2 음성을 결합한 결합 음성을 생성하는 단계 및 결합 텍스트 및 결합 음성의 학습 결합 시 에러 레이트(error rate)가 설정된 기준 레이트(reference rate)보다 작으면 결합 텍스트 및 결합 음성을 상기 음성 합성 모델에 추가하는 단계를 포함하는 음성 합성 시스템의 동작방법을 제공한다.

명세서

발명의 명칭: 음성 합성 시스템의 동작방법

기술분야

- [1] 본 발명은 음성 합성 시스템의 동작방법에 관한 것으로서, 더욱 상세하게는, 장문의 대상 텍스트에 대응하는 대상 합성 음성을 출력할 수 있는 음성 합성 시스템의 동작방법에 관한 것이다.
- [2] 나아가, 본 발명은 2021년 01월 13일 출원된 한국특허출원 제10-2021-0004856호의 이익을 주장하며, 그 내용 전부는 본 명세서에 포함된다.

배경기술

- [3] 인공지능(Artificial Intelligence, AI)은 인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해 능력 등을 컴퓨터 프로그램으로 실현한 기술을 의미한다.
- [4] 현재 개발되고 있는 인공지능은 대화형 사용자 인터페이스(Conversational User Interface, CUI)를 구현하기 위해 필요한 기술들에 집중되어 있다. 그러한 기술로 음성인식(STT), 자연어 이해(NLU), 자연어 생성(NLG), 음성합성(TTS)이 있다.
- [5] 음성합성 기술은 인공지능을 통한 대화형 사용자 인터페이스 구현을 위한 핵심 기술로서, 인간이 발화하는 것과 같은 소리를 컴퓨터나 기계를 통하여 만들어내는 것이다.
- [6] 기존의 음성합성은 고정 합성 단위(Fixed Length Unit)인 단어, 음절, 음소를 조합하여 파형을 만들어내는 방식(1세대), 말뭉치를 이용한 가변 합성 단위 연결방식(2세대)에서, 3세대 모델로 발전하였다. 3세대 모델은 음성인식을 위한 음향모델링에 주로 사용하는 HMM(Hidden Markov Model)방식을 음성합성에 적용하여, 적절한 크기의 데이터베이스를 이용한 고품질 음성합성을 구현하였다.
- [7] 기존의 음성합성이 특정 화자의 음색과 억양, 말투를 학습하기 위해서는 그 화자의 음성 데이터가 최소 5시간, 고품질의 음성 출력을 위해서는 10시간 이상 필요했다. 하지만 같은 사람의 음성 데이터를 그만큼 확보하는 것은 많은 비용과 시간이 소요되었다.
- [8] 또한, 모델의 학습이 완료되고 음성을 출력하려고 할 때 학습된 텍스트의 길이보다 더 긴 길이의 텍스트가 입력되게 되면 해당 텍스트에 따른 음성을 합성하는데 오류가 발생하는 문제가 발생하게 되었다.
- [9] 최근들어, 학습한 텍스트보다 긴 텍스트가 입력되는 경우에도 오류가 발생하지 않고 음성을 합성하기 위한 방법을 연구하고 있다.

발명의 상세한 설명

기술적 과제

- [10] 본 발명의 목적은, 장문의 대상 텍스트에 대응하는 대상 합성 음성을 출력할 수 있는 음성 합성 시스템의 동작방법을 제공함에 있다.

- [11] 본 발명의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 발명의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 발명의 목적 및 장점들은 특허 청구 범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제 해결 수단

- [12] 본 발명에 따른 음성 합성 시스템의 동작방법은, 제1 텍스트와 상기 제1 텍스트에 대한 제1 음성 및 제2 텍스트와 상기 제2 텍스트에 대한 제2 음성이 입력되는 단계, 상기 제1, 2 텍스트 및 상기 제1, 2 음성을 커리큘럼 러닝(Curriculum learning)에 적용하여 학습한 음성 합성 모델을 생성하는 단계 및 음성 출력을 위한 대상 텍스트 입력 시, 상기 음성 합성 모델을 기반으로 상기 대상 텍스트에 대응하는 대상 합성 음성을 출력하는 단계를 포함하고, 상기 음성 합성 모델을 생성하는 단계는, 상기 제1, 2 텍스트를 결합한 결합 텍스트 및 상기 제1, 2 음성을 결합한 결합 음성을 생성하는 단계 및 상기 결합 텍스트 및 상기 결합 음성의 학습 결합 시 에러 레이트(error rate)가 설정된 기준 레이트(reference rate)보다 작으면 상기 결합 텍스트 및 상기 결합 음성을 상기 음성 합성 모델에 추가하는 단계를 포함할 수 있다.
- [13] 상기 결합 텍스트는, 상기 제1, 2 텍스트 및 상기 제1, 2 텍스트를 구분하는 텍스트 토큰(text token)을 포함할 수 있다.
- [14] 상기 결합 음성은, 상기 제1, 2 음성 및 상기 제1, 2 음성을 구분하는 멜스펙트로그램 토큰(mel spectrogram-token)을 포함할 수 있다.
- [15] 상기 텍스트 토큰 및 상기 멜스펙트로그램 토큰은, 1초 내지 2초의 시간 구간을 갖을 수 있다.
- [16] 상기 텍스트 토큰 및 상기 멜스펙트로그램 토큰은, 묵음 구간일 수 있다.
- [17] 상기 음성 합성 모델에 추가하는 단계는, 상기 텍스트 토큰 및 상기 멜스펙트로그램 토큰을 기준으로 결합할 수 있다.
- [18] 상기 음성 합성 모델에 추가하는 단계 이전에, 상기 결합 텍스트 및 상기 결합 음성의 학습 결합 시 배치 사이즈(batch size)가 설정된 기준 배치 사이즈보다 작으면 상기 결합 텍스트 및 상기 결합 음성을 초기화하는 단계를 더 포함할 수 있다.
- [19] 상기 음성 합성 모델에 추가하는 단계는, 상기 에러 레이트가 상기 기준 레이트보다 크면 상기 결합 텍스트 및 상기 결합 음성을 초기화할 수 있다.

발명의 효과

- [20] 본 발명에 따른 음성 합성 시스템의 동작방법은, 커리큘럼 러닝으로 학습하여 단문 텍스트로부터 긴 장문 텍스트까지의 음성 합성 모델을 생성하여, 장문의 텍스트에 대한 음성 합성이 용이한 이점이 있다.
- [21] 또한, 본 발명에 따른 음성 합성 시스템의 동작방법은, 장문 텍스트에 대한 음성

합성 시 자연스러운 음성을 출력할 수 있는 이점이 있다.

- [22] 한편, 본 발명의 효과는 이상에서 언급한 효과들로 제한되지 않으며, 이하에서 설명할 내용으로부터 통상의 기술자에게 자명한 범위 내에서 다양한 효과들이 포함될 수 있다.

도면의 간단한 설명

- [23] 도 1은 본 발명에 따른 음성 합성 시스템의 제어 구성을 나타낸 제어블록도이다.
- [24] 도 2는 본 발명에 따른 결합 텍스트 및 결합 음성을 설명하기 위한 간략도이다.
- [25] 도 3 및 도 4는 본 발명에 따른 음성 합성 시스템에 대한 실험 결과를 나타낸 도 및 표이다.
- [26] 도 5는 본 발명에 따른 음성 합성 시스템의 동작방법을 나타낸 순서도이다.

발명의 실시를 위한 형태

- [27] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.
- [28] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수개의 관련된 기재된 항목들의 조합 또는 복수개의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [29] 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.
- [30] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수개의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 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.
- [31] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [32] 이하, 본 발명에 따른 바람직한 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [33] 도 1은 본 발명에 따른 음성 합성 시스템의 제어 구성을 나타낸 제어블록도 및 도 2는 본 발명에 따른 결합 텍스트 및 결합 음성을 설명하기 위한 간략도이다.
- [34] 도 1 및 도 2를 참조하면, 음성 합성 시스템(100)은 인코더(110), 어텐션(120) 및 디코더(130)를 포함할 수 있다.
- [35] 인코더(110)는 입력된 텍스트를 처리하고, 디코더(130)는 음성을 출력할 수 있다.
- [36] 여기서 인코더(110)는 텍스트를 벡터로 압축할 수 있으며, 디코더(130)는 어텐션(120)에서 출력된 음성을 출력할 수 있다. 이때, 음성은 멜스펙트로그램(Mel spectrogram)일 수 있다.
- [37] 이때, 인코더(110)는 텍스트를 텍스트 정보를 표현하는 숫자 벡터로 변환할 수 있으며, 이에 한정을 두지 않는다.
- [38] 어텐션(120)은 텍스트로부터 음성을 생성할 수 있다.
- [39] 어텐션(120)은 어텐션 매커니즘을 활용하여 입력 시퀀스의 기울기 소실 현상을 최소화할 수 있다. 어텐션 매커니즘은 다음과 같은 함수 형태로 표현된다.
- [40] $\text{Attention}(Q, K, V) = \text{Attention Value}$
- [41] 여기서, Q는 쿼리(Query)를 뜻하며 이는 해당 시점에 디코더(130)에서의 은닉 상태를 반영하는 값이다. K는 키(Key), V는 벨류(Value)를 뜻하며 이는 모든 시점의 인코더(110)의 은닉 상태들을 반영한다. 어텐션 함수는 키에 대한 쿼리의 유사도를 구한 후 이를 벨류에 반영하여 중요한 부분을 강조한다.
- [42] 디코더(130) 현재 시점의 값을 반영하는 쿼리와 인코더 값을 반영하는 키를 dot product 해준 뒤 softmax를 취해주면 쿼리가 인코더(110)의 어떤 부분에 집중해야 예측값에 도움이 될지 가중치 형태로 표현된다. 이때 각각의 가중치를 어텐션 가중치(Attention Weight)라 한다. 이후, 어텐션의 최종 결과값을 얻기 위해서 인코더(110)의 값을 반영하고 있는 벨류와 어텐션 가중치 값들을 곱해주면 어텐션 값(Attention Value)이 나오게 된다. 어텐션 값은 인코더(110)의 문맥을 포함하고 있기 때문에 컨텍스트 벡터(context vector)라고 불리기도 한다.
- [43] 최종적으로 컨텍스트 벡터를 디코더(130) 현재 시점의 은닉상태와 결합(concatenate)하여 출력 시퀀스를 구하게 된다.
- [44] 어텐션(130)은 텍스트 및 음성을 합성하여 음성 합성 모델을 생성할 수 있다.

- [45] 즉, 어텐션(130)는 학습 데이터가 입력되면, 학습 데이터를 설정된 커리큘럼 러닝(Curriculum learning)에 적용하여 학습한 음성 합성 모델을 생성하고, 음성 출력을 위한 대상 텍스트가 입력되면 상기 음성 합성 모델을 기반으로 대상 텍스트에 대응하는 대상 합성 음성을 생성할 수 있다.
- [46] 실시 예에서, 학습 데이터는 제1 텍스트와 상기 제1 텍스트에 대한 제1 음성 및 제2 텍스트와 상기 제2 텍스트에 대한 제2 음성으로 이루어진 것으로 설명하지만, 텍스트의 개수 및 음성의 개수에 대하여 한정을 두지 않는다.
- [47] 제1 텍스트 및 제1 음성이 입력되는 경우, 어텐션(130)은 인코더(110)로부터 제1 텍스트에 대한 벡터를 제공받으며, 디코더(120)로부터 제1 음성에 대한 멜스펙트로그램을 제공받을 수 있다.
- [48] 또한, 제2 텍스트 및 제2 음성이 입력되는 경우, 어텐션(130)은 인코더(110)로부터 제2 텍스트에 대한 벡터를 제공받으며, 디코더(120)로부터 제2 음성에 대한 멜스펙트로그램을 제공받을 수 있다.
- [49] 음성 합성 모델을 생성하기 위해, 어텐션(130)은 제1, 2 텍스트 및 제1, 2 음성을 설정된 커리큘럼 러닝(Curriculum learning)에 적용하여 학습할 수 있다.
- [50] 먼저, 어텐션(130)은 제1, 2 텍스트 및 제1, 2 음성을 개별 학습 처리할 수 있다.
- [51] 이후, 어텐션(130)은 제1, 2 텍스트를 결합한 결합 텍스트 및 제1, 2 음성을 결합한 결합 음성을 생성할 수 있다.
- [52] 여기서, 결합 텍스트는 제1, 2 텍스트 및 제1, 2 텍스트를 구분하기 위한 텍스트 토큰(text token)을 포함할 수 있으며, 결합 음성은 제1, 2 음성 및 제1, 2 음성을 구분하기 위한 멜스펙트로그램 토큰(mel spectrogram-token)을 포함할 수 있다.
- [53] 여기서, 텍스트 토큰 및 멜스펙트로그램 토큰은 1초 내지 2초의 시간 구간이며, 묵음 구간일 수 있다.
- [54] 텍스트 토큰 및 멜스펙트로그램 토큰은 두 개의 제1, 2 텍스트가 서로 자연스럽게 연결되어, 하나의 텍스트로 인식할 수 있도록 하기 위한 도구이다.
- [55] 실시예에서, 어텐션(130)은 2개의 제1, 2 텍스트 및 2개의 제1, 2 음성으로 하나의 결합 텍스트 및 결합 음성을 생성하여 학습하는 것으로 설명하였으나, 복수의 텍스트 및 복수의 음성이 존재하는 경우 결합 텍스트 및 결합 음성도 복수 개로 생성할 수 있으며, 이에 한정을 두지 않는다.
- [56] 어텐션(130)은 결합 텍스트 및 결합 음성의 학습 결합 시 배치 사이즈(batch size)가 설정된 기준 배치 사이즈보다 작으면 결합 텍스트 및 결합 음성을 초기화할 수 있다.
- [57] 결합 텍스트 및 결합 음성을 초기화하는 경우, 어텐션(130)은 제1, 2 텍스트 및 제1, 2 음성을 재결합하거나, 결합 텍스트 및 결합 음성을 생성하지 않을 수 있다.
- [58] 이후, 어텐션(130)은 결합 텍스트 및 결합 음성의 학습 결합 시 에러 레이트(error rate)가 설정된 기준 레이트(reference rate)보다 작으면 결합 텍스트 및 결합 음성을 음성 합성 모델에 추가할 수 있다.
- [59] 또한, 어텐션(130)은 에러 레이트가 기준 레이트보다 크면 결합 텍스트 및 결합

음성을 초기화할 수 있다.

[60] 어텐션(130)은 음성 출력을 위한 대상 텍스트 입력 시, 상기 음성 합성 모델을 기반으로 상기 대상 텍스트에 대응하는 대상 합성 음성을 출력할 수 있다.

[61] 이때, 어텐션(130)은 인코더(110)로부터 대상 텍스트에 대한 벡터가 입력되면, 디코더(120)로 대상 합성 음성이 출력되게 할 수 있다.

[62] 도 3 및 도 4는 본 발명에 따른 음성 합성 시스템에 대한 실험 결과를 나타낸 도 및 표이다.

[63] 도 3 및 도 4를 참조하면, 음성 합성 시스템(100)은 기본 모델로 Tacotron2를 사용하고, 커리큘럼 러닝(curriculum learning)을 위한 모델을 추가 적용하여 음성 합성을 수행할 수 있다.

[64] 여기서, 배치 사이즈(Batch size)는 기본적으로 12를 사용했으며, 한정된 GPU 용량 내에서 커리큘럼 러닝(curriculum learning)을 통해 장문의 문장을 합성하기 위해 n개의 문장을 합쳐서 학습할 경우 자동으로 배치 사이즈가 1/n의 크기로 줄어들도록 설정하였다.

[65] 음성 합성 시스템(100)은 긴 문장을 한 번에 합성할 수 있는 장점이 있다.

[66] 이를 검증하기 위해, 음성 합성 시스템(100)에는 소설 Harry Potter 책의 script를 사용하여 음성합성 테스트를 진행하였으며, 합성된 음성의 길이와 시간에 따라 평가를 진행하였다.

[67] 도 3은 기존 모델과 합성할 수 있는 문장의 길이를 비교하기 위해 content based attention을 사용한 Tacotron 모델과 location sensitive attention 을 사용한 Tacotron2 모델을 사용하여 제안하는 모델과 비교하였다.

[68] 또한, 제안하는 모델에서 curriculum learning의 적용 여부에 따른 모델의 성능 비교 실험을 진행하였다. Length robustness 실험을 위해 Google Cloud Speech-To-Text를 통해 합성된 음성을 transcript로 변환하고 이를 원본과 비교해 character error rate(CER) 을 측정하였다.

[69] 도 3에서 우리가 제안한 모델은 5분 30초 길이의 음성(약 4400개의 characters)을 합성할 때까지 CER 10% 초반을 넘기지 않는 반면, content based attention model은 10초대에서, location sensitive attention model은 30초대에서 20%가 넘어가는 것을 확인할 수 있었다.

[70] 커리큘럼 러닝(curriculum learning)을 적용하지 않았을 때 (DLTTS1)와 두 문장만을 이어 붙여 이를 적용했을 때(DLTTS2) 5분 이상의 문장 합성 환경에서 CER이 60% 대까지 올라가지만 세 문장을 이어붙여 커리큘럼 러닝(curriculum learning)을 적용했을 때 (DLTTS3) CER이 10% 초반대로 떨어지는 것을 확인할 수 있었다.

[71] 문서 단위의 음성을 합성할 때 인코더(110)와 디코더(120) 간의 어텐션(130)이 제대로 형성되지 않으면 합성된 음성에서 word repeating이나 word skipping같은 어텐션 에러(attention error)가 발생하게 된다.

[72] 여기서, 도 4를 참조하면, 본 발명의 음성 합성 시스템(100)에서 제공하는

모델은 기존 모델에 비해 document-level에서 굉장히 낮은 어텐션 에러 레이트(attention error rate)를 보인다는 것을 알 수 있다.

- [73] 문장의 길이별로 임의의 200개 document를 test하여 어텐션 에러가 발생하는 횟수를 측정하였다. 실험 결과 content based attention을 사용한 Tacotron모델은 30초 이상의 문장을 합성할 때 높은 어텐션 에러 레이트(attention error rate)를 보이며, location sensitive attention을 사용한 Tacotron2모델은 합성된 문장의 길이가 1분 이상일 때 높은 어텐션 에러 레이트(attention error rate)를 보였다.
- [74] 반면 본 발명의 음성 합성 시스템(100)에서 Document-level Neural TTS model은 5분 이상의 문장을 합성할 때에도 비교적 낮은 어텐션 에러 레이트(attention error rate)를 보였다. 이는 document-level의 문장도 안정적으로 합성할 수 있음을 보여준다. 또한 커리큘럼 러닝(curriculum learning)을 사용하지 않았을 때 5분 이상의 문장 환경에서 어텐션 에러 레이트(attention error rate)가 50%가 넘어간 반면, 2문장으로 커리큘럼 러닝(curriculum learning)을 실행했을 때에는 25%, 3문장으로 실행했을 때에는 1% 정도의 어텐션 에러 레이트(attention error rate)가 측정되는 것을 알 수 있었다. 이를 통해 우리는 문서 단위의 음성합성을 할 때 커리큘럼 러닝(curriculum learning)이 필수적인 요소임을 알 수 있었다.
- [75] 도 5는 본 발명에 따른 음성 합성 시스템의 동작방법을 나타낸 순서도이다.
- [76] 도 5를 참조하면, 음성 합성 시스템(100)은 제1 텍스트와 상기 제1 텍스트에 대한 제1 음성 및 제2 텍스트와 상기 제2 텍스트에 대한 제2 음성이 입력될 수 있다(S110).
- [77] 음성 합성 시스템(100)은 상기 제1, 2 텍스트 및 상기 제1, 2 음성을 커리큘럼 러닝(Curriculum learning)에 적용하여 상기 제1, 2 텍스트를 결합한 결합 텍스트 및 상기 제1, 2 음성을 결합한 결합 음성을 생성할 수 있다(S120).
- [78] 음성 합성 시스템(100)은 상기 결합 텍스트 및 상기 결합 음성의 학습 결합 시 배치 사이즈(batch size)가 설정된 기준 배치 사이즈보다 작은지 판단하고(S130), 기준 배치 사이즈 보다 작으면 상기 결합 텍스트 및 상기 결합 음성을 초기화할 수 있다(S140).
- [79] 기준 배치 사이즈보다 큰 경우, 음성 합성 시스템(100)은 상기 결합 텍스트 및 상기 결합 음성의 학습 결합 시 에러 레이트(error rate)가 설정된 기준 레이트(reference rate)보다 작은지 판단하고(S150), 기준 레이트보다 작으면 상기 결합 텍스트 및 상기 결합 음성을 상기 음성 합성 모델에 생성 및 추가할 수 있다(S160).
- [80] (S150) 단계 이후, 기준 레이트보다 큰 경우, 음성 합성 시스템(100)은 상기 결합 텍스트 및 상기 결합 음성을 초기화할 수 있다(S170).
- [81] (S160) 단계 이후, 음성 합성 시스템(100)은 음성 출력을 위한 대상 텍스트 입력 시, 상기 음성 합성 모델을 기반으로 상기 대상 텍스트에 대응하는 대상 합성 음성을 출력할 수 있다(S180)
- [82] 이상에서 실시 예들에 설명된 특징, 구조, 효과 등은 본 발명의 적어도 하나의

실시 예에 포함되며, 반드시 하나의 실시 예에만 한정되는 것은 아니다. 나아가, 각 실시 예에서 예시된 특징, 구조, 효과 등은 실시 예들이 속하는 분야의 통상의 지식을 가지는 자에 의해 다른 실시 예들에 대해서도 조합 또는 변형되어 실시 가능하다. 따라서 이러한 조합과 변형에 관계된 내용들은 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

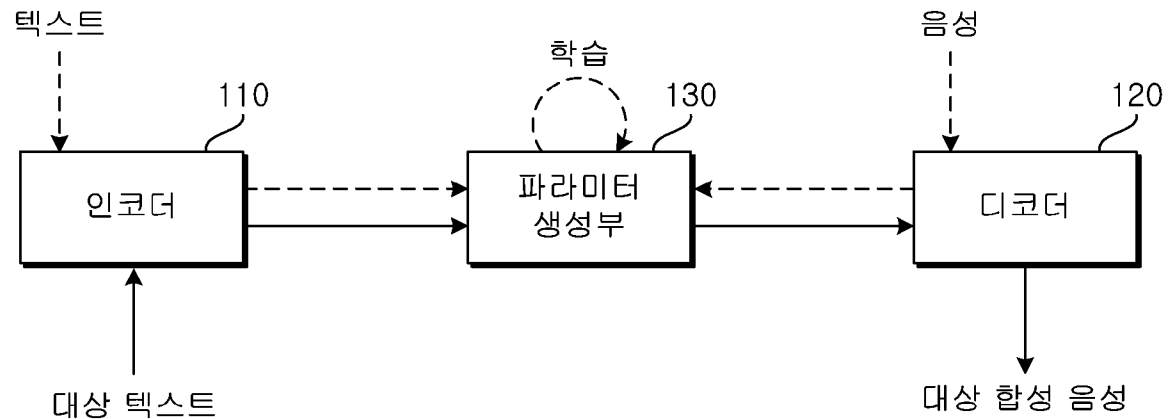
- [83] 또한, 이상에서 실시 예를 중심으로 설명하였으나 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 실시 예의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러가지의 변형과 응용이 가능함을 알 수 있을 것이다. 예를 들어, 실시 예에 구체적으로 나타난 각 구성 요소는 변형하여 실시할 수 있는 것이다. 그리고 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

청구범위

- [청구항 1] 제1 텍스트와 상기 제1 텍스트에 대한 제1 음성 및 제2 텍스트와 상기 제2 텍스트에 대한 제2 음성이 입력되는 단계;
상기 제1, 2 텍스트 및 상기 제1, 2 음성을 커리큘럼 러닝(Curriculum learning)에 적용하여 학습한 음성 합성 모델을 생성하는 단계; 및
음성 출력을 위한 대상 텍스트 입력 시, 상기 음성 합성 모델을 기반으로 상기 대상 텍스트에 대응하는 대상 합성 음성을 출력하는 단계를 포함하고,
상기 음성 합성 모델을 생성하는 단계는,
상기 제1, 2 텍스트를 결합한 결합 텍스트 및 상기 제1, 2 음성을 결합한 결합 음성을 생성하는 단계; 및
상기 결합 텍스트 및 상기 결합 음성의 학습 결합 시 에러 레이트(error rate)가 설정된 기준 레이트(reference rate)보다 작으면 상기 결합 텍스트 및 상기 결합 음성을 상기 음성 합성 모델에 추가하는 단계를 포함하는, 음성 합성 시스템의 동작방법.
- [청구항 2] 제 1 항에 있어서,
상기 결합 텍스트는,
상기 제1, 2 텍스트 및 상기 제1, 2 텍스트를 구분하는 텍스트 토큰(text token)을 포함하는,
음성 합성 시스템의 동작방법.
- [청구항 3] 제 2 항에 있어서,
상기 결합 음성은,
상기 제1, 2 음성 및 상기 제1, 2 음성을 구분하는 멜스펙트로그램 토큰(mel spectrogram-token)을 포함하는,
음성 합성 시스템의 동작방법.
- [청구항 4] 제 3 항에 있어서,
상기 텍스트 토큰 및 상기 멜스펙트로그램 토큰은,
1초 내지 2초의 시간 구간을 갖는,
음성 합성 시스템의 동작방법.
- [청구항 5] 제 3 항에 있어서,
상기 텍스트 토큰 및 상기 멜스펙트로그램 토큰은,
목음 구간인,
음성 합성 시스템의 동작방법.
- [청구항 6] 제 3 항에 있어서,
상기 음성 합성 모델에 추가하는 단계는,
상기 텍스트 토큰 및 상기 멜스펙트로그램 토큰을 기준으로 결합하는,
음성 합성 시스템의 동작방법.

- [청구항 7] 제 1 항에 있어서,
상기 음성 합성 모델에 추가하는 단계 이전에,
상기 결합 텍스트 및 상기 결합 음성의 학습 결합 시 배치 사이즈(batch size)가 설정된 기준 배치 사이즈보다 작으면 상기 결합 텍스트 및 상기 결합 음성을 초기화하는 단계를 더 포함하는,
음성 합성 시스템의 동작방법.
- [청구항 8] 제 1 항에 있어서,
상기 음성 합성 모델에 추가하는 단계는,
상기 에러 레이트가 상기 기준 레이트보다 크면 상기 결합 텍스트 및 상기 결합 음성을 초기화하는,
음성 합성 시스템의 동작방법.

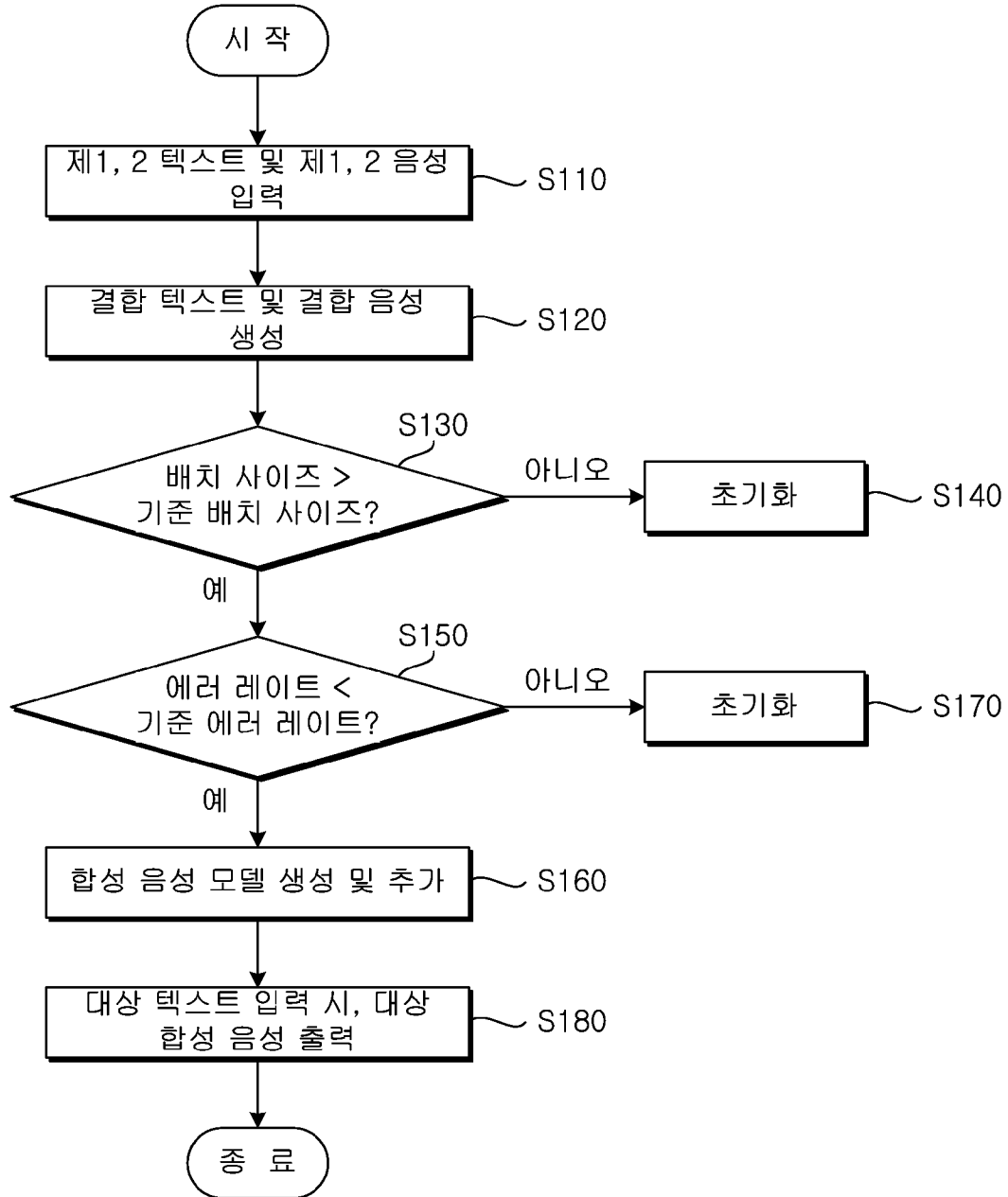
[도1]

100

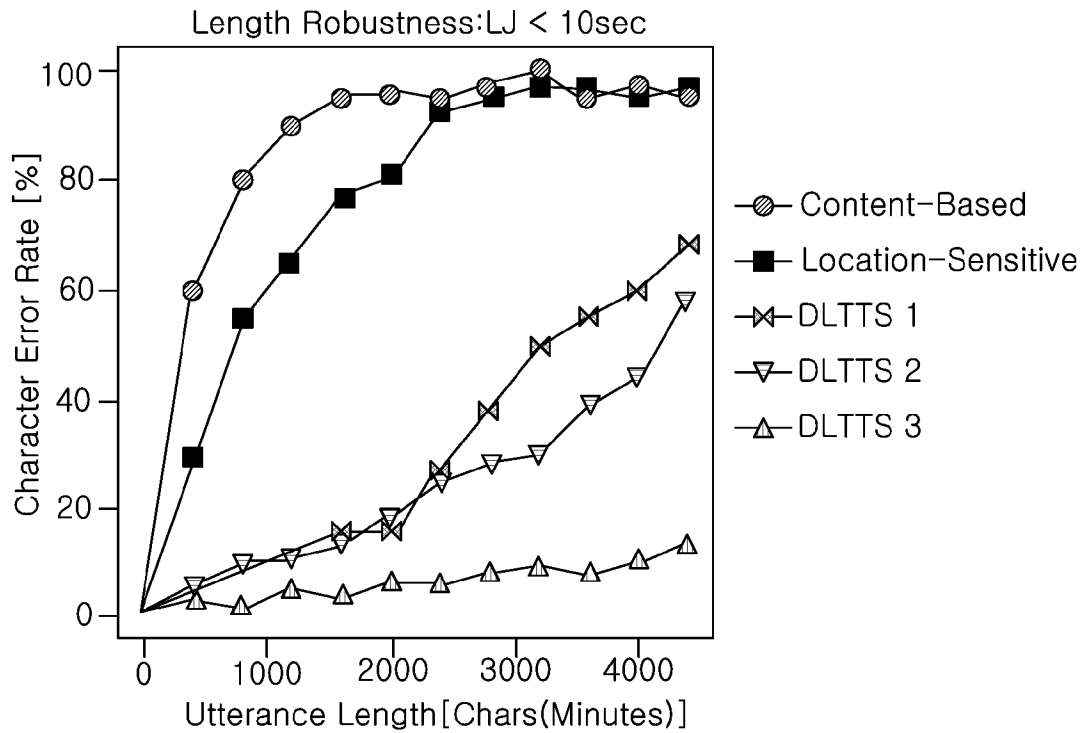
[도2]



[도3]



[도4]



[도5]

Method	Text length	Attention error rate
Content-Based	10sec	4/200
	30sec	200/200
	1min	200/200
Location-Sensitive	10sec	0/200
	30sec	24/200
	1min	189/200
Document-Level Neural TTS (curriculum learning with 1 sentence)	10sec	0/200
	1min	1/200
	3min	44/200
	5min	101/200
Document-Level Neural TTS (curriculum learning with 2 sentence)	10sec	0/200
	1min	0/200
	3min	8/200
	5min	57/200
Document-Level Neural TTS (curriculum learning with 3 sentence)	10sec	0/200
	1min	0/200
	3min	0/200
	5min	2/200

INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2021/095116

A. CLASSIFICATION OF SUBJECT MATTER G10L 13/08(2006.01)i; G10L 13/02(2006.01)i; G10L 25/18(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 13/08(2006.01); G10L 13/033(2013.01); G10L 13/06(2006.01); G10L 15/16(2006.01); 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: 텍스트(text), 커리큘럼 러닝(curriculum learning), 텍스트-음성 합성(text-to-speech synthesis), 에러 레이트(error rate), 토큰(token)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	HWANG, Sung-Woong et al. Document-Level Neural TTS Using Curriculum Learning and Attention Masking. IEEE Access. Vol. 9, pages 8954-8960, 04 January 2021. See pages 8955-8956 and 8960.	1-8
Y	KR 10-2019-0085882 A (NEOSAPIENCE, INC.) 19 July 2019 (2019-07-19) See paragraph [0131]; and claim 1.	1-8
A	JP 2020-126141 A (NIPPON TELEGR & TELEPH CORP.) 20 August 2020 (2020-08-20) See claims 1-7.	1-8
A	JP 2010-237323 A (TOSHIBA CORP.) 21 October 2010 (2010-10-21) See claims 1-10.	1-8
A	CONG, Yahuan et al. PPSpeech: Phrase based Parallel End-to-End TTS System. arXiv: 2008.02490, August 2020 [Retrieved on 12 April 2022]. Retrieved from <https://arxiv.org/abs/2008.02490>. See sections 2-3.	1-8
☐ 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 14 April 2022		Date of mailing of the international search report 15 April 2022
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex-Daejeon Building 4, 189 Cheongsaro, Seo-gu, Daejeon 35208 Facsimile No. +82-42-481-8578		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2021/095116

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
KR	10-2019-0085882	A	19 July 2019	CN	111587455	A	25 August 2020
				EP	3739572	A1	18 November 2020
				EP	3739572	A4	08 September 2021
				JP	2021-511533	A	06 May 2021
				US	2020-0082807	A1	12 March 2020
JP	2020-126141	A	20 August 2020	JP	7036054	B2	15 March 2022
				WO	2020-162190	A1	13 August 2020
JP	2010-237323	A	21 October 2010	JP	5457706	B2	02 April 2014
				US	2012-0065961	A1	15 March 2012
				WO	2010-116549	A1	14 October 2010

A. 발명이 속하는 기술분류(국제특허분류(IPC)) G10L 13/08(2006.01)i; G10L 13/02(2006.01)i; G10L 25/18(2013.01)i		
B. 조사된 분야 조사된 최소문헌(국제특허분류를 기재) G10L 13/08(2006.01); G10L 13/033(2013.01); G10L 13/06(2006.01); G10L 15/16(2006.01); 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: 텍스트(text), 커리큘럼 러닝(curriculum learning), 텍스트-음성 합성(text-to-speech synthesis), 에러 레이트(error rate), 토큰(token)		
C. 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	SUNG-WOONG HWANG 등, 'Document-Level Neural TTS Using Curriculum Learning and Attention Masking', IEEE Access, Vol. 9, pages 8954-8960, 2021.01.04 페이지 8955-8956, 8960	1-8
Y	KR 10-2019-0085882 A (네오사피엔스 주식회사) 2019.07.19 단락 [0131]; 및 청구항 1	1-8
A	JP 2020-126141 A (NIPPON TELEGR & TELEPH CORP.) 2020.08.20 청구항 1-7	1-8
A	JP 2010-237323 A (TOSHIBA CORP.) 2010.10.21 청구항 1-10	1-8
A	YAHUAN CONG 등, 'PPSpeech: Phrase based Parallel End-to-End TTS System', arXiv: 2008.02490, 2020.08 [검색일: 2022.04.12]. 출처 <https://arxiv.org/abs/2008.02490> 섹션 2-3	1-8
☐ 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 2022년04월14일(14.04.2022)		국제조사보고서 발송일 2022년04월15일(15.04.2022)
ISA/KR의 명칭 및 우편주소 대한민국 특허청 (35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사) 팩스 번호 +82-42-481-8578		심사관 고재용 전화번호 +82-42-481-8212

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2019-0085882 A	2019/07/19	CN 111587455 A	2020/08/25
		EP 3739572 A1	2020/11/18
		EP 3739572 A4	2021/09/08
		JP 2021-511533 A	2021/05/06
		US 2020-0082807 A1	2020/03/12
JP 2020-126141 A	2020/08/20	JP 7036054 B2	2022/03/15
		WO 2020-162190 A1	2020/08/13
JP 2010-237323 A	2010/10/21	JP 5457706 B2	2014/04/02
		US 2012-0065961 A1	2012/03/15
		WO 2010-116549 A1	2010/10/14