(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0010259
(43) 공개일자 2022년01월25일

(51) 국제특허분류(Int. Cl.)
G10L 21/02 (2006.01) *G10L 15/06* (2006.01)
G10L 15/22 (2006.01) *G10L 15/26* (2006.01)
G10L 25/87 (2013.01)
(52) CPC특허분류
G10L 21/02 (2021.08)
G10L 15/22 (2013.01)
(21) 출원번호 10-2020-0088919
(22) 출원일자 2020년07월17일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
강태균
경기도 수원시 영통구 동탄원천로881번길 35, 508동 704호 (매탄동, 주공그린빌)
(74) 대리인
특허법인 무한

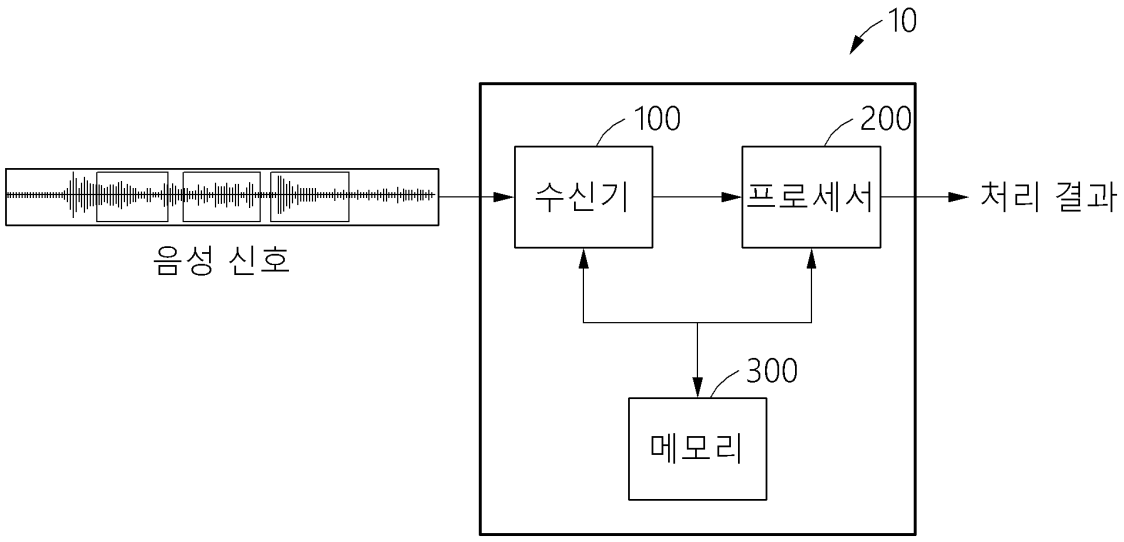
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **음성 신호 처리 방법 및 장치**

(57) 요 약

음성 신호 처리 방법 및 장치가 개시된다. 일 실시예에 따른 음성 신호 처리 방법은, 음성 신호에 기초한 입력 토큰을 수신하는 단계와, 상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하는 단계와, 상기 제1 확률 값들의 순위에 기초하여 상기 제1 확률 값들 중 적어도 하나의 값을 수정하는 단계와, 수정된 확률 값에 기초하여 상기 음성 신호를 처리하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 15/26 (2013.01)
G10L 25/87 (2013.01)
G10L 2015/0635 (2013.01)

명 세 서

청구범위

청구항 1

음성 신호에 기초한 입력 토큰을 수신하는 단계;

상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하는 단계;

상기 제1 확률 값들의 순위에 기초하여 상기 제1 확률 값들 중 적어도 하나의 값을 수정하는 단계; 및

수정된 확률 값에 기초하여 상기 음성 신호를 처리하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 2

제1항에 있어서,

상기 수정하는 단계는,

상기 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단하는 단계; 및

판단 결과에 기초하여 상기 제1 확률 값을 수정하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 3

제2항에 있어서,

상기 제1 후보 출력 토큰은,

문장 끝(end of sentence)에 대응되는 토큰인

음성 신호 처리 방법.

청구항 4

제2항에 있어서,

상기 판단 결과에 기초하여 상기 제1 확률 값을 수정하는 단계는,

상기 제1 확률 값이 상기 미리 결정된 순위에 포함되지 않을 경우,

상기 제1 확률 값을 감소시키는 단계

를 포함하는 음성 신호 처리 방법.

청구항 5

제4항에 있어서,

상기 감소시키는 단계는,

상기 제1 확률 값의 대수 값을 음의 무한대로 수정하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 6

제1항에 있어서,

상기 처리하는 단계는,

상기 음성 신호에 기초하여 텍스트를 출력하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 7

제1항에 있어서,

상기 처리하는 단계는,

상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제2 확률 값들을 계산하는 단계; 및

상기 제1 확률 값들 및 상기 제2 확률 값들에 기초하여 출력 토큰을 결정하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 8

제7항에 있어서,

상기 결정하는 단계는,

상기 제1 확률 값들 및 제2 확률 값들 각각을 가중합하는 단계; 및

가중합한 값이 가장 큰 후보 출력 토큰을 상기 출력 토큰으로 결정하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 9

음성 신호에 기초한 입력 토큰을 수신하는 수신기; 및

상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하고, 상기 제1 확률 값들의 순위에 기초하여 상기 제1 확률 값들 중 적어도 하나의 값을 수정하고, 수정된 확률 값에 기초하여 상기 음성 신호를 처리하는 프로세서

를 포함하는 음성 신호 처리 장치.

청구항 10

제9항에 있어서,

상기 프로세서는,

상기 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단하고,

판단 결과에 기초하여 상기 제1 확률 값을 수정하는

를 포함하는 음성 신호 처리 장치.

청구항 11

제10항에 있어서,

상기 제1 후보 출력 토큰은,

문장 끝(end of sentence)에 대응되는 토큰인

음성 신호 처리 장치.

청구항 12

제10항에 있어서,

상기 프로세서는,

상기 제1 확률 값이 상기 미리 결정된 순위에 포함되지 않을 경우,

상기 제1 확률 값을 감소시키는

음성 신호 처리 장치.

청구항 13

제12항에 있어서,

상기 프로세서는,

상기 제1 확률 값의 대수 값을 음의 무한대로 수정하는

음성 신호 처리 장치.

청구항 14

제9항에 있어서,

상기 프로세서는,

상기 음성 신호에 기초하여 텍스트를 출력하는

음성 신호 처리 장치.

청구항 15

제9항에 있어서,

상기 프로세서는,

상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제2 확률 값들을 계산하고,

상기 제1 확률 값들 및 상기 제2 확률 값들에 기초하여 출력 토큰을 결정하는

음성 신호 처리 장치.

청구항 16

제15항에 있어서,

상기 프로세서는,

상기 제1 확률 값들 및 제2 확률 값들 각각을 가중합하고,

가중합한 값이 가장 큰 후보 출력 토큰을 상기 출력 토큰으로 결정하는

음성 신호 처리 장치.

청구항 17

음성 신호에 기초한 입력 토큰을 수신하는 단계;

상기 입력 토큰에 기초하여 복수의 제1 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하는 단계;

상기 입력 토큰에 기초하여 복수의 제2 후보 출력 토큰에 대응하는 제2 확률 값들을 계산하는 단계;

상기 제1 확률 값들 및 상기 제2 확률 값들 중 적어도 하나에 기초하여 비언어적 토큰 및 상기 비언어적 토큰에 대응하는 확률을 생성하는 단계; 및

상기 비언어적 토큰에 대응하는 확률에 기초하여 상기 음성 신호를 처리하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 18

제17항에 있어서,

상기 생성하는 단계는,

상기 복수의 제1 후보 출력 토큰에 포함된 상기 비언어적 토큰을 상기 복수의 제2 후보 출력 토큰에 복사하는 단계; 및

상기 제1 확률 값들에 포함된 상기 비언어적 토큰에 대응하는 확률을 상기 제2 확률 값들에 복사하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 19

제17항에 있어서,

상기 생성하는 단계는,

상기 복수의 제1 후보 출력 토큰에 포함된 상기 비언어적 토큰을 상기 제2 후보 출력 토큰에 복사하는 단계; 및

상기 제2 확률 값들 중에서 가장 큰 값을 복사된 비언어적 토큰에 대응하는 확률로 매핑하는 단계

를 포함하는 음성 신호 처리 방법.

청구항 20

제17항에 있어서,

상기 비언어적 토큰이 상기 복수의 제2 후보 출력 토큰에 등록되어 있는지 여부를 판단하는 단계

를 더 포함하는 음성 신호 처리 방법.

발명의 설명

기 술 분 야

[0001] 아래 실시예들은 음성 신호 처리 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 신경 언어 모델(neural language model)의 경우 텍스트만으로 학습될 수 있다는 점이 장점이나, 텍스트만으로 학습되기 때문에 음성 신호를 보지 못하여 필연적으로 발생하는 미스매치(mismatch) 문제를 가진다.

[0003] 가장 흔한 문제는 삭제 오류(deletion error)로, 언어 모델은 오디오 정보가 없기 때문에 문법적으로 하나의 문장이 완성되는 부분에서 문장의 끝으로, 확률이 높아져 음성인식이 실제 문장보다 더 빨리 종료되는 문제점을 가진다.

[0004] 또한, 전사 규칙에 따라 기침 소리나 흠(hmm) 소리와 같이, 음성 인식 모델에 존재하는 토큰이라도 언어 모델에 존재하지 않는 경우가 하지 않는 경우가 있는데, 이러한 음성 신호에 대하여 언어 모델은 적절한 확률을 계산할 수 없다는 문제점을 가진다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0005] 일 실시예에 따른 음성 신호 처리 방법은, 음성 신호에 기초한 입력 토큰을 수신하는 단계와, 상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하는 단계와, 상기 제1 확률 값들의 순위에 기초하여 상기 제1 확률 값들 중 적어도 하나의 값을 수정하는 단계와, 수정된 확률 값에 기초하여 상기 음성 신호를 처리하는 단계를 포함한다.

[0006] 상기 수정하는 단계는, 상기 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단하는 단계와, 판단 결과에 기초하여 상기 제1 확률 값을 수정하는 단계를 포함할 수 있다.

[0007] 상기 제1 후보 출력 토큰은, 문장 끝(end of sentence)에 대응되는 토큰일 수 있다.

[0008] 상기 판단 결과에 기초하여 상기 제1 확률 값을 수정하는 단계는, 상기 제1 확률 값이 상기 미리 결정된 순위에 포함되지 않을 경우, 상기 제1 확률 값을 감소시키는 단계를 포함할 수 있다.

[0009] 상기 감소시키는 단계는, 상기 제1 확률 값의 대수 값을 음의 무한대로 수정하는 단계를 포함할 수 있다.

[0010] 상기 처리하는 단계는, 상기 음성 신호에 기초하여 텍스트를 출력하는 단계를 포함할 수 있다.

[0011] 상기 처리하는 단계는, 제2 뉴럴 네트워크에서 상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제2 확률 값들을 계산하는 단계와, 상기 제1 확률 값들 및 상기 제2 확률 값들에 기초하여 출력 토큰을 결정하는 단계를 포함할 수 있다.

[0012] 상기 결정하는 단계는, 상기 제1 확률 값들 및 제2 확률 값들 각각을 가중합하는 단계와, 가중합한 값이 가장 큰 후보 출력 토큰을 상기 출력 토큰으로 결정하는 단계를 포함할 수 있다.

[0013] 일 실시예에 따른 음성 신호 처리 장치는, 음성 신호에 기초한 입력 토큰을 수신하는 수신기와, 상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하고, 상기 제1 확률 값들의 순위에 기초하여 상기 제1 확률 값들 중 적어도 하나의 값을 수정하고, 수정된 확률 값에 기초하여 상기 음성 신호를 처리하는 프로세서를 포함한다.

[0014] 상기 프로세서는, 상기 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단하고, 판단 결과에 기초하여 상기 제1 확률 값을 수정할 수 있다.

[0015] 상기 제1 후보 출력 토큰은, 문장 끝(end of sentence)에 대응되는 토큰일 수 있다.

[0016] 상기 프로세서는, 상기 제1 확률 값이 상기 미리 결정된 순위에 포함되지 않을 경우, 상기 제1 확률 값을 감소시킬 수 있다.

[0017] 상기 프로세서는, 상기 제1 확률 값의 대수 값을 음의 무한대로 수정할 수 있다.

[0018] 상기 프로세서는, 상기 음성 신호에 기초하여 텍스트를 출력할 수 있다.

[0019] 상기 프로세서는, 상기 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제2 확률 값들을 계산하고, 상기 제1 확률 값들 및 상기 제2 확률 값들에 기초하여 출력 토큰을 결정할 수 있다.

[0020] 상기 프로세서는, 상기 제1 확률 값들 및 제2 확률 값들 각각을 가중합하고, 가중합한 값이 가장 큰 후보 출력 토큰을 상기 출력 토큰으로 결정할 수 있다.

[0021] 다른 실시예에 따른 음성 신호 처리 방법은, 음성 신호에 기초한 입력 토큰을 수신하는 단계와, 상기 입력 토큰에 기초하여 복수의 제1 후보 출력 토큰에 대응하는 제1 확률 값들을 계산하는 단계와, 상기 입력 토큰에 기초하여 복수의 제2 후보 출력 토큰에 대응하는 제2 확률 값들을 계산하는 단계와, 상기 제1 확률 값들 및 상기 제2 확률 값들 중 적어도 하나에 기초하여 비언어적 토큰 및 상기 비언어적 토큰에 대응하는 확률을 생성하는 단계와, 상기 비언어적 토큰에 대응하는 확률에 기초하여 상기 음성 신호를 처리하는 단계를 포함한다.

[0022] 상기 생성하는 단계는, 상기 복수의 제1 후보 출력 토큰에 포함된 상기 비언어적 토큰을 상기 복수의 제2 후보 출력 토큰에 복사하는 단계와, 상기 제1 확률 값들에 포함된 상기 비언어적 토큰에 대응하는 확률을 상기 제2 확률 값들에 복사하는 단계를 포함할 수 있다.

[0023] 상기 생성하는 단계는, 상기 복수의 제1 후보 출력 토큰에 포함된 상기 비언어적 토큰을 상기 제2 후보 출력 토큰에 복사하는 단계와, 상기 제2 확률 값들 중에서 가장 큰 값을 복사된 비언어적 토큰에 대응하는 확률로 매핑하는 단계를 포함할 수 있다.

[0024] 상기 음성 신호 처리 방법은, 상기 비언어적 토큰이 상기 복수의 제2 후보 출력 토큰에 등록되어 있는지 여부를 판단하는 단계를 더 포함할 수 있다.

도면의 간단한 설명

[0025] 도 1은 일 실시예에 따른 음성 신호 처리 장치의 개략적인 블록도를 나타낸다.

도 2는 도 1에 도시된 음성 신호 처리 장치가 음성 신호를 처리하는 개략적인 과정을 도시한다.

도 3은 도 1에 도시된 음성 신호 처리 장치가 후보 출력 토큰에 대응하는 확률을 수정하는 동작의 예를 나타낸다.

도 4는 도 1에 도시된 음성 신호 처리 장치가 비언어적 토큰에 대응하는 확률을 생성하는 동작의 예를 나타낸다.

도 5는 도 1에 도시된 음성 신호 처리 장치의 동작의 순서도를 나타낸다.

도 6은 음성 신호 처리 방법의 예를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0026] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다. 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.

[0027] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "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.

[0028] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속

하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0029] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0030] 또한, 실시 예의 구성 요소를 설명하는 데 있어서, 제 1, 제 2, A, B, (a), (b) 등의 용어를 사용할 수 있다. 이러한 용어는 그 구성 요소를 다른 구성 요소와 구별하기 위한 것일 뿐, 그 용어에 의해 해당 구성 요소의 본질이나 차례 또는 순서 등이 한정되지 않는다. 어떤 구성 요소가 다른 구성요소에 "연결", "결합" 또는 "접속"된다고 기재된 경우, 그 구성 요소는 그 다른 구성요소에 직접적으로 연결되거나 접속될 수 있지만, 각 구성 요소 사이에 또 다른 구성 요소가 "연결", "결합" 또는 "접속"될 수도 있다고 이해되어야 할 것이다.

[0031] 어느 하나의 실시 예에 포함된 구성요소와, 공통적인 기능을 포함하는 구성요소는, 다른 실시 예에서 동일한 명칭을 사용하여 설명하기로 한다. 반대되는 기재가 없는 이상, 어느 하나의 실시 예에 기재한 설명은 다른 실시 예에도 적용될 수 있으며, 중복되는 범위에서 구체적인 설명은 생략하기로 한다.

[0033] 도 1은 일 실시예에 따른 음성 신호 처리 장치의 개략적인 블록도를 나타낸다.

[0034] 도 1을 참조하면, 음성 신호 처리 장치(10)는 음성 신호를 수신하여 처리하고, 처리 결과를 출력할 수 있다. 예를 들어, 음성 신호 처리 장치(10)는 음성 신호를 인식하거나 음성 신호에 대하여 번역을 수행할 수 있다. 이 때, 처리 결과는 텍스트 또는 음성을 포함할 수 있다.

[0035] 음성 신호 처리 장치(10)는 음성 신호를 처리하여 음성 인식 성능을 향상시킬 수 있다. 음성 신호 처리 장치(10)는 뉴럴 네트워크를 이용하여 적어도 하나의 토큰에 대응하는 확률을 수정함으로써 음성 인식의 성능을 향상시킬 수 있다.

[0036] 음성 신호 처리 장치(10)는 엔드-투-엔드 ASR(Automatic Speech Recognition) 방식으로 음성 신호를 처리할 수 있다.

[0037] 예를 들어, 엔드-투-엔드 ASR 방식은 시퀀스-투-시퀀스(sequence-to-sequence) 방식의 음성 신호처리 방법을 포함할 수 있다. 엔드-투-엔드 ASR 방식에 대해서는 도 2를 참조하여 자세하게 설명한다.

[0038] 음성 신호 처리 장치(10)는 뉴럴 네트워크를 이용하여 음성 신호를 처리할 수 있다. 뉴럴 네트워크(또는 인공 신경망)는 기계학습과 인지과학에서 생물학의 신경을 모방한 통계학적 학습 알고리즘을 포함할 수 있다. 뉴럴 네트워크는 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 의미할 수 있다.

[0039] 뉴럴 네트워크는 심층 뉴럴 네트워크 (Deep Neural Network)를 포함할 수 있다. 뉴럴 네트워크는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), 퍼셉트론(perceptron), FF(Feed Forward), RBF(Radial Basis Network), DFF(Deep Feed Forward), LSTM(Long Short Term Memory), GRU(Gated Recurrent Unit), AE(Auto Encoder), VAE(Variational Auto Encoder), DAE(Denoising Auto Encoder), SAE(Sparse Auto Encoder), MC(Markov Chain), HN(Hopfield Network), BM(Boltzmann Machine), RBM(Restricted Boltzmann Machine), DBN(Depp Belief Network), DCN(Deep Convolutional Network), DN(Deconvolutional Network), DCIGN(Deep Convolutional Inverse Graphics Network), GAN(Generative Adversarial Network), LSM(Liquid State Machine), ELM(Extreme Learning Machine), ESN(Echo State Network), DRN(Deep Residual Network), DNC(Differentiable Neural Computer), NTM(Neural Turning Machine), CN(Capsule Network), KN(Kohonen Network) 및 AN(Attention Network)를 포함할 수 있다.

[0040] 음성 신호 처리 장치(10)는 음성 인식뿐만 아니라 시퀀스-투-시퀀스(sequence-to-sequence) 모델의 앙상블(ensemble)에서 모델이 수신하는 입력의 차이 때문에 일부 토큰의 미스매치를 수정하는 다양한 분야에 활용될 수 있다.

[0041] 예를 들어, 음성 신호 처리 장치(10)는 번역 알고리즘, 또는 이미지 캡션(image caption) 알고리즘과 언어 모델

(language model)이 함께 사용되는 경우 및 앙상블 모델(ensemble model)을 활용한 텍스트 생성(text generation) 등에서 사용될 수 있다.

[0042] 음성 신호 처리 장치(10)는 수신기(100) 및 프로세서(200)를 포함한다. 음성 신호 처리 장치(10)는 메모리(300)를 더 포함할 수 있다.

[0043] 수신기(100)는 음성 신호에 기초한 입력 토큰을 수신할 수 있다. 수신기(100)는 마이크로폰과 같은 하드웨어 또는 수신 인터페이스를 포함할 수 있다. 수신기(100)는 음성 신호를 수신하여 입력 토큰을 생성할 수 있다.

[0044] 토큰은 자연어로 이루어진 말뭉치(corpus)를 토큰화(tokenization)한 것을 의미할 수 있다. 수신기(100)는 말뭉치를 특정 기준에 기초하여 토큰화할 수 있다.

[0045] 수신기(100)는 의미를 가지는 문자의 집합을 기준으로 말뭉치를 토큰화할 수 있다. 예를 들어, 수신기(100)는 음소, 음절 또는 단어를 기준으로 말뭉치를 토큰화할 수 있다.

[0046] 수신기(100)는 토큰의 시퀀스를 수신할 수 있다. 토큰의 시퀀스는 하나 이상의 시간 대에 수신되는 토큰의 집합을 의미할 수 있다. 수신기(100)는 입력 토큰 및/또는 입력 토큰의 시퀀스를 프로세서(200)로 출력할 수 있다.

[0047] 프로세서(200)는 메모리(300)에 저장된 데이터를 처리할 수 있다. 프로세서(200)는 메모리(300)에 저장된 컴퓨터로 읽을 수 있는 코드(예를 들어, 소프트웨어) 및 프로세서(200)에 의해 유발된 인스트럭션(instruction)들을 실행할 수 있다.

[0048] "프로세서(200)"는 목적하는 동작들(desired operations)을 실행시키기 위한 물리적인 구조를 갖는 회로를 가지는 하드웨어로 구현된 데이터 처리 장치일 수 있다. 예를 들어, 목적하는 동작들은 프로그램에 포함된 코드(code) 또는 인스트럭션들(instructions)을 포함할 수 있다.

[0049] 예를 들어, 하드웨어로 구현된 데이터 처리 장치는 마이크로프로세서(microprocessor), 중앙 처리 장치(central processing unit), 프로세서 코어(processor core), 멀티-코어 프로세서(multi-core processor), 멀티프로세서(multiprocessor), ASIC(Application-Specific Integrated Circuit), FPGA(Field Programmable Gate Array)를 포함할 수 있다.

[0050] 프로세서(200)는 뉴럴 네트워크를 이용하여 입력 토큰을 처리할 수 있다. 프로세서(200)는 뉴럴 네트워크를 학습시킬 수 있다. 프로세서(200)는 학습된 뉴럴 네트워크를 이용하여 입력 토큰을 처리할 수 있다.

[0051] 프로세서(200)는 수신한 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산할 수 있다. 후보 출력 토큰은 입력 토큰의 다음에 출력될 토큰을 의미할 수 있다. 후보 출력 토큰은 그에 대응하는 제1 확률 값을 가질 수 있다. 후보 출력 토큰에 대응하는 확률은 후보 출력 토큰이 출력 토큰이 될 확률을 의미할 수 있다.

[0052] 프로세서(200)는 후보 출력 토큰에 대응하는 제1 확률 값에 기초하여 출력 토큰을 결정할 수 있다. 프로세서(200)는 인코더 및 디코더 구조를 갖는 뉴럴 네트워크를 이용하여 제1 확률 값을 계산할 수 있다. 프로세서(200)가 제1 확률 값들을 계산하는 과정은 도 2를 참조하여 자세하게 설명한다.

[0053] 프로세서(200)는 제1 확률 값들의 순위에 기초하여 제1 확률 값들 중 적어도 하나의 값을 수정할 수 있다. 순위는 확률 값들의 크기의 순위를 의미할 수 있다. 예를 들어, 순위는 확률 값의 대수 값의 크기의 순위를 의미할 수 있다.

[0054] 확률 값이 큰 경우 높은 순위를 가질 수 있고, 확률 값이 작은 경우 낮은 순위를 가질 수 있다. 순위의 숫자가 작을수록 높은 순위를 의미할 수 있다. 다시 말해, 1 순위가 가장 높은 순위를 의미할 수 있다. 확률 값이 가장 큰 값이 1 순위 확률을 가질 수 있다.

[0055] 프로세서(200)는 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단할 수 있다. 예를 들어, 제1 후보 출력 토큰은 문장 끝(end of sentence)에 대응되는 토큰일 수 있다. 이하에서, 문장 끝에 대응되는 토큰을 EOS 토큰이라고 지칭한다.

[0056] 프로세서(200)는 판단 결과에 기초하여 제1 확률 값을 수정할 수 있다. 프로세서(200)는 제1 확률 값이 미리 결정된 순위에 포함되지 않을 경우, 제1 확률 값을 감소시킬 수 있다. 반대로, 프로세서(200)는 제1 확률 값이 미리 결정된 순위 내에 포함될 경우, 제1 확률 값을 변화시키지 않고 그대로 출력 토큰을 결정할 수 있다.

[0057] 제1 확률 값이 미리 결정된 순위에 포함된 다는 것은 제1 확률 값의 크기의 순위가 미리 결정된 순위보다 높거나 같다는 것을 의미한다. 제1 확률 값이 미리 결정된 순위에 포함되지 않는다는 것은 제1 확률 값의 크기의 순위가 미리 결정된 순위보다 낮다는 것을 의미한다.

[0058] 다시 말해, 제1 확률 값이 미리 결정된 순위에 포함된다는 것은 제1 확률 값의 크기가 미리 결정된 순위에 대응하는 확률 값보다 크거나 같다는 것을 의미하고, 제1 확률 값이 미리 결정된 순위에 포함되지 않는다는 것은 제1 확률 값의 크기가 미리 결정된 순위에 대응하는 확률 값보다 작다는 것을 의미할 수 있다.

[0059] 프로세서(200)는 제1 확률 값이 미리 결정된 순위 내에 포함될 경우, 제1 확률 값의 대수 값을 음의 무한대로 수정할 수 있다. 다시 말해, 프로세서(200)는 제1 확률 값을 0으로 수정할 수 있다.

[0060] 프로세서(200)는 수정된 확률 값에 기초하여 음성 신호를 처리할 수 있다. 예를 들어, 프로세서(200)는 수정된 확률 값에 기초하여 음성 신호에 기초하여 텍스트를 출력할 수 있다. 여기서, 텍스트는 음성 신호의 인식 결과 또는 번역 결과를 포함할 수 있다.

[0061] 프로세서(200)는 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제2 확률 값들을 계산할 수 있다. 이하에서, 복수의 제1 후보 출력 토큰에 대응하는 확률을 제1 확률이라 지칭하고, 복수의 제2 후보 출력 토큰에 대응하는 확률을 제2 확률이라고 지칭한다.

[0062] 다시 말해, 프로세서(200)는 입력 토큰에 기초하여 복수의 제1 후보 출력 토큰에 대응하는 제1 확률 값들을 계산할 수 있고, 입력 토큰에 기초하여 복수의 제2 후보 출력 토큰에 대응하는 제2 확률 값들을 계산할 수 잇다.

[0063] 프로세서(200)는 계산한 제1 확률 값들 및 제2 확률 값들 중 적어도 하나에 기초하여 비언어적 토큰 및 비언어적 토큰에 대응하는 확률을 생성할 수 있다. 비언어적 토큰은 자연어는 아니지만 음성 신호에 포함될 수 있는 신호에 대응하는 토큰을 의미할 수 있다. 비언어적 토큰은 넌스피치 버벌 사운드(non-speech verbal sound) 또는 어쿠스틱 이벤트(acoustic event)에 대응하는 토큰을 포함할 수 있다. 예를 들어, 비언어적 토큰은 기침(cough), 침묵(silence), 감탄사(exclamation)(예를 들어, 흠(hmm)) 등의 음성 신호에 대응되는 토큰을 포함할 수 있다.

[0064] 프로세서(200)는 비언어적 토큰이 복수의 제2 후보 출력 토큰에 등록되어 있는지 여부를 판단할 수 있다. 프로세서(200)는 복수의 제2 후보 출력 토큰 중에 비언어적 토큰이 등록되어 있지 않은 경우, 제2 후보 출력 토큰에 비언어적 토큰을 생성할 수 있다.

[0065] 프로세서(200)는 복사 또는 매핑을 통해 비언어적 토큰 및 비언어적 토큰에 대응하는 확률을 생성할 수 있다. 프로세서(200)는 복수의 제1 후보 출력 토큰에 포함된 비언어적 토큰을 복수의 제2 후보 출력 토큰에 복사할 수 있다.

[0066] 프로세서(200)는 제1 확률 값들에 포함된 비언어적 토큰에 대응하는 확률을 제2 확률 값들에 복사할 수 있다. 또는, 프로세서(200)는 제2 확률 값들 중에서 가장 큰 값을 복사된 비언어적 토큰에 대응하는 확률로 매핑할 수 있다.

[0067] 실시예에 따라, 복사 및 매핑 이외에 다양한 방법으로 제2 후보 출력 토큰에 대응하는 확률을 생성할 수 있다.

[0068] 프로세서(200)는 제1 확률 값들 및 제2 확률 값들에 기초하여 출력 토큰을 결정할 수 있다.

[0069] 제1 확률 값들의 계산과 제2 확률 값들의 계산은 상이한 뉴럴 네트워크를 이용하여 수행될 수 있다. 제1 확률 값들 및 제2 확률 값들의 계산은 도 3을 참조하여 자세하게 설명한다.

[0070] 프로세서(200)는 제1 확률 값들 및 제2 확률 값들 각각을 가중합할 수 있다. 프로세서(200)는 가중합한 값이 가장 큰 후보 출력 토큰을 출력 토큰으로 결정할 수 있다. 프로세서(200)는 동일한 후보 출력 토큰에 대응하는 제1 확률 값들 및 제2 확률 값들을 가중합할 수 있다.

[0071] 출력 토큰은 연속되는 토큰 시퀀스에서 입력 토큰 다음에 위치할 토큰을 의미할 수 있다. 프로세서(200)는 출력 토큰을 지속적으로 추정함으로써 토큰 시퀀스로 구성되는 문장을 출력할 수 있다.

[0072] 출력 토큰을 결정하는 과정은 도 3을참조하여 상세하게 설명한다.

[0073] 메모리(300)는 프로세서(200)에 의해 실행가능한 인스트럭션들(또는 프로그램)을 저장할 수 있다. 예를 들어, 인스트럭션들은 프로세서(200)의 동작 및/또는 프로세서(200)의 각 구성의 동작을 실행하기 위한 인스트럭션들을 포함할 수 있다.

[0074] 메모리(300)는 휘발성 메모리 장치 또는 불휘발성 메모리 장치로 구현될 수 있다.

[0075] 휘발성 메모리 장치는 DRAM(dynamic random access memory), SRAM(static random access memory), T-RAM(thyristor RAM), Z-RAM(zero capacitor RAM), 또는 TTRAM(Twin Transistor RAM)으로 구현될 수 있다.

[0076] 불휘발성 메모리 장치는 EEPROM(Electrically Erasable Programmable Read-Only Memory), 플래시(flash) 메모리, MRAM(Magnetic RAM), 스핀전달토크 MRAM(Spin-Transfer Torque(STT)-MRAM), Conductive Bridging RAM(CBRAM), FeRAM(Ferroelectric RAM), PRAM(Phase change RAM), 저항 메모리(Resistive RAM(RRAM)), 나노 튜브 RRAM(Nanotube RRAM), 폴리머 RAM(Polymer RAM(PoRAM)), 나노 부유 게이트 메모리(Nano Floating Gate Memory(NFGM)), 홀로그래픽 메모리(holographic memory), 분자 전자 메모리 소자(Molecular Eelectronic Memory Device), 또는 절연 저항 변화 메모리(Insulator Resistance Change Memory)로 구현될 수 있다.

[0078] 도 2는 도 1에 도시된 음성 신호 처리 장치가 음성 신호를 처리하는 개략적인 과정을 도시한다.

[0079] 도 2를 참조하면, 프로세서(200)는 뉴럴 네트워크를 이용하여 음성 신호를 처리할 수 있다. 프로세서(200)가 이용하는 뉴럴 네트워크는 인코더(210) 및 디코더(230)를 포함할 수 있다.

[0080] 디코더(230)는 인코딩된 특징을 입력으로 수신할 수 있다. 예를 들어, 디코더(230)는 인공신경망 내 디코더(230)의 전단에 연결된 인코더(210)로부터 입력을 수신할 수 있다.

[0081] 인코더(210) 및 디코더(230)는 DNN(Deep Neural Network), RNN(Recurrent Neural Network) 또는 RDNN(Recurrent Deep Neural Network) 등으로 구현될 수 있다. 인코더(210) 및 디코더(230) 내의 레이어들의 노드들은 비선형적으로 서로 영향을 주는 관계일 수 있으며, 각 노드들로부터 출력되는 값들, 노드들 사이의 관계들 등 뉴럴 네트워크의 파라미터들은 학습에 의해 최적화될 수 있다.

[0082] 인코더(210) 및 디코더(230)는 통합된 네트워크 구조로서, 입력 토큰의 시퀀스로부터 인식 결과의 시퀀스를 생성할 수 있다. 예를 들어, 시퀀스 투 시퀀스 구조로 구현된 인코더와 디코더(230)는 입력된 음성 신호로부터 입력된 음성 신호에 대응하는 인식 결과를 직접 생성할 수 있다. 인코더(210)와 디코더(230)는 입력 시퀀스로부터 인식 결과의 시퀀스를 생성하도록 미리 학습될 수 있다.

[0083] 인코더(210)는 음성 신호를 수신하여 분석할 수 있고, 디코더(230)는 매 스텝 마다 기존의 히스토리(history)의 다음에 연결된 워드피스(wordpiece)를 추정할 수 있다.

[0084] 디코더(230)는 전체 음성 신호에 대응하는 문장이 모두 추정된 후에는 EOS 토큰을 출력할 수 있다. 프로세서(200)는 EOS 토큰이 주어진 경우 디코딩을 완료하고 최종 음성 인식 결과를 사용자에게 반환할 수 있다.

[0085] 인코더는 입력 토큰 시퀀스의를 인코딩하여 인코딩된 특징을 생성할 수 있다. 인코더는 입력 토큰의 시퀀스로부터 특징을 추출하여, 인코딩된 정보를 생성할 수 있다. 인코딩된 특징은 디코더(230)로 인가될 수 있다. 디코더(230)는 인코딩된 특징에 기초하여 인식 결과를 생성할 수 있다.

[0086] 인코딩된 특징을 수신한 디코더(230)는 입력 토큰인 시작 토큰에 기초하여, 복수의 후보 출력 토큰 중 하나의 토큰을 출력 토큰으로 결정할 수 있다. 디코더(230)는 결정된 출력 토큰을 다음 입력 토큰으로 결정할 수 있다.

[0087] 디코더(230)는 매 타임 스텝에서 인코더(210)로부터 계산된 정보를 바탕으로 출력 토큰을 구하는데, 이때 이전 스텝까지 선택되었던 입력 토큰들에 종속되어 출력 토큰을 구할 수 있다.

[0088] 예를 들어, 결정한 출력 토큰을 새로운 입력 토큰으로 결정한 디코더(230)는 결정한 출력 토큰에 기초하여 다음 후보 출력 토큰의 후보들의 확률들을 예측할 수 있다.

[0090] 도 3은 도 1에 도시된 음성 신호 처리 장치가 후보 출력 토큰에 대응하는 확률을 수정하는 동작의 예를 나타낸다.

[0091] 도 3을 참조하면, 프로세서(200)는 음성과 쌍(pair)로 이루어진 학습 말뭉치(training corpus)에 포함되지 않는 다양한 텍스트들에 대한 인식 성능을 향상시키기 위해 신경 언어 모델(neural language model)을 포함하는 제2 뉴럴 네트워크를 이용할 수 있다.

[0092] 제1 뉴럴 네트워크 외부의 언어 모델을 포함하는 제2 뉴럴 네트워크는 제1 뉴럴 네트워크에서 사용하는 말뭉치 중에서 자주 등장하지 않는 단어나 컨텍스트 바이어스(context bias) 등과 같이, 실제 음성과 텍스트의 쌍이 많지 않아서 제1 뉴럴 네트워크가 독립적으로 학습하기 어려운 케이스를 인식하는 것을 도울 수 있다.

[0093] 프로세서(200)는 다양한 방식으로 제1 뉴럴 네트워크의 출력과 제2 뉴럴 네트워크의 출력을 결합할 수 있다. 예를 들어, 프로세서(200)는 쉘로우-퓨전(shallow-fusion) 방식을 이용하여 제1 뉴럴 네트워크와 제2 뉴럴 네트워크의 출력을 가중합함으로써 두 네트워크의 출력을 결합할 수 있다.

[0094] 이 때, 프로세서(200)는 제1 뉴럴 네트워크와 제2 뉴럴 네트워크 중 적어도 하나의 네트워크가 계산한 확률 값을 수정함으로써 음성 신호의 처리 성능을 향상시킬 수 있다.

[0095] 이하에서, 예시적으로, 제1 뉴럴 네트워크와 제2 뉴럴 네트워크의 출력을 결합하는 동작을 설명한다.

[0096] 실시예에 따라, 제1 뉴럴 네트워크와 제2 뉴럴 네트워크는 다양한 뉴럴 네트워크 모델을 포함할 수 있다.

[0097] 이하의 예시에서 편의상 제1 뉴럴 네트워크는 음성 인식 모델을 포함하고, 제2 뉴럴 네트워크는 언어 모델을 포함하는 것으로 설명하였지만, 이와 반대로, 제1 뉴럴 네트워크가 언어 모델을 포함하고, 제2 뉴럴 네트워크가 음성 인식 모델을 포함할 수도 있다. 다시 말해, 제1 뉴럴 네트워크와 제2 뉴럴 네트워크 중 하나의 뉴럴 네트워크는 음성 인식 모델을 포함하고, 다른 뉴럴 네트워크는 언어 모델을 포함할 수 있다.프로세서(200)는 제2 뉴럴 네트워크로 외부 언어 모델을 사용하는 음성 인식 모델(예를 들어, 엔드-투-엔드 음성 인식 모델)에서 특정 토큰(제1 후보 출력 토큰)에 대응하는 확률을 제1 뉴럴 네트워크의 확률 값에 의존하여 결정할 수 있다.

[0098] 또는, 프로세서(200)는 제2 뉴럴 네트워크의 확률 값을 수정함으로써 제1 뉴럴 네트워크와 제2 뉴럴 네트워크를 이용하여 계산되는 특정 토큰에 대응하는 최종 확률 값을 수정할 수 있다.

[0099] 프로세서(200)는 음성 인식 모델에서 컨텍스트가 주어졌을 때, 각 후보 출력 토큰에 대한 대수 우도(log likelihood)를 계산하고, 그 결과를 정렬하여 특정 토큰(제1 후보 토큰)이 미리 결정된 순위 내에 포함되지 않을 경우, 특정 토큰에 대응하는 확률 값을 수정하여 특정 토큰이 최종 빔(beam)에 포함되지 수정할 수 있다.

[0100] 제1 확률의 계산과 제2 확률의 계산은 서로 상이한 뉴럴 네트워크를 이용하여 수행될 수 있다. 프로세서(200)는 제1 뉴럴 네트워크를 이용하여 제1 확률을 계산하고, 제2 뉴럴 네트워크를 이용하여 제2 확률을 계산할 수 있다.

[0101] 제1 뉴럴 네트워크 및 제2 뉴럴 네트워크는 인코더(210) 및 디코더(230) 중 적어도 하나를 포함할 수 있다. 제1 뉴럴 네트워크 및 제2 뉴럴 네트워크가 포함하는 인코더(210) 및 디코더(230)의 동작은 도 2에서 설명한 것과 동일할 수 있다.

[0102] 프로세서(200)는 제1 뉴럴 네트워크 및 제2 뉴럴 네트워크를 서로 상이한 데이터 셋을 이용하여 학습시킬 수 있다. 예를 들어, 프로세서(200)는 제1 뉴럴 네트워크를 음성 및 텍스트가 결합된 데이터 셋을 이용하여 학습시킬 수 있다. 프로세서(200)는 제2 뉴럴 네트워크를 텍스트로 이루어진 데이터 셋을 이용하여 학습시킬 수 있다.

[0103] 프로세서(200)는 제1 확률과 제2 확률에 기초하여 출력 토큰을 결정할 수 있다.

[0104] 예를 들어, 제1 뉴럴 네트워크는 음성 인식 모델(예를 들어, 엔드-투-엔드 ASR)을 포함할 수 있고, 제2 뉴럴 네트워크는 언어 모델(Language Model(LM))을 포함할 수 있다. 또는, 위에서 설명한 것과 같이, 제1 뉴럴 네트워크가 언어 모델을 포함하고, 제2 뉴럴 네트워크가 음성 인식 모델을 포함할 수도 있다.

[0105] 제1 뉴럴 네트워크는 도 2에서 설명한 인코더(210) 및 디코더(230)를 이용하여 복수의 후보 출력 토큰에 대응하는 확률 값들을 계산할 수 있다.

[0106] 제1 뉴럴 네트워크는 음성을 인코더(210)의 입력으로 이용하여 인식결과인 텍스트(예를 들어, 단어)의 시퀀스를 출력하는 디코더(230)로 구성될 수 있다.

[0107] 필요에 따라, 제2 뉴럴 네트워크로 단어의 연속이 얼마나 문장에서의 확률이 높은지 판단하는 디코더(미도시)만으로 이루어진 언어 모델을 활용하여 음성 신호 처리의 성능을 높일 수 있다.

[0108] 프로세서(200)는 음성 신호로부터 음성 특징을 추출할 수 있다. 입력된 음성 신호는 복수의 프레임들 별로 정보를 포함하는 음성 신호이고, 음성 특징은 적어도 하나의 프레임 단위로 추출된 정보의 시퀀스일 수 있고, 다차원의 벡터로 표현될 수 있다.

[0109] 프로세서(200)는 제1 뉴럴 네트워크에 포함된 디코더(230)와 제2 뉴럴 네트워크에 포함된 디코더(미도시)의 앙상블(ensemble)을 이용하여 입력 음성의 시퀀스로부터 인식 결과의 시퀀스를 생성할 수 있다.

[0110] 제1 뉴럴 네트워크에 포함된 디코더(230)와 제2 뉴럴 네트워크에 포함된 디코더는 토큰 단위로 각각의 인식 결과를 출력할 수 있고, 각각의 인식 결과를 앙상블 가중치에 따라 앙상블하여 최종 인식 결과를 생성할 수 있다.

[0111] 예를 들어, 제1 뉴럴 네트워크에 포함된 디코더(230)는 입력 음성 신호와 이전에 결정된 인식 결과에 기초하여 복수의 후보 출력 토큰을 결정할 수 있고, 제2 뉴럴 네트워크에 포함된 디코더는 이전에 결정된 인식 결과에 기초하여 복수의 후보 출력을 결정할 수 있으며, 프로세서(200)는 각각의 후보 출력 토큰들을 앙상블 가중치에 따라 앙상블하여 최종 인식 결과를 생성할 수 있다.

[0112] 인코더(210)와 디코더(230)는 입력된 음성 신호에 대응하는 정답 텍스트 쌍의 시퀀스로부터 인식 결과의 시퀀스를 생성하도록 미리 학습될 수 있고, 제2 뉴럴 네트워크의 디코더는 임의의 텍스트 시퀀스로부터 인식 결과의 시퀀스를 생성하도록 미리 학습될 수 있다.

[0113] 인코더(210)는 음성 특징을 인코딩하여 인코딩된 특징을 생성할 수 있다. 인코더(210)는 음성 특징의 차원(dimension)을 변환시켜, 인코딩된 정보를 생성할 수 있다.

[0114] 인코딩된 특징은 제1 뉴럴 네트워크의 디코더(230)로 인가될 수 있다. 디코더(230)는 토큰 단위로, 인코딩된 특징과 이전에 결정된 인식 결과 기초하여 복수의 후보 출력 토큰을 생성하고, 제2 뉴럴 네트워크의 디코더는 토큰 단위로, 이전에 결정된 인식 결과에 기초하여 출력 토큰의 후보들을 생성할 수 있다.

[0115] 프로세서(200)는 제1 뉴럴 네트워크의 디코더(230)에 의한 인식 결과 및 제2 뉴럴 네트워크의 디코더에 의한 인식 결과를 미리 정해진 앙상블 가중치에 따라 앙상블하여 최종 인식 결과를 생성할 수 있다.

[0116] 도 3의 예시에서, 제1 확률 값들(310)은 제1 뉴럴 네트워크에 의해 계산된 확률을 의미할 수 있다. 제2 확률 값들(350)은 제2 뉴럴 네트워크에 의해 계산된 확률을 의미할 수 있다.

[0117] 프로세서(200)는 복수의 제1 후보 출력 토큰 및 그에 대응하는 제1확률 값들(310)을 정렬(sort)할 수 있다. 이후, 프로세서(200)는 제1 확률 값들(310)의 순위에 기초하여 제1 확률 값들 중 적어도 하나의 값을 수정할 수 있다.

[0118] 프로세서(200)는 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단할 수 있다.

[0119] 도 3의 예시에서, 제1 확률 값들은 복수의 후보 토큰에 대응하는 확률 값들을 의미할 수 있다. 도 3의 예시에 표시된 확률 값들은 확률 값의 대수 값을 의미할 수 있다.

[0120] 예를 들어, 후보 출력 토큰 from에 대응하는 확률 값의 대수 값은 -0.67이고, 후보 출력 토큰 at에 대응하는 확률 값의 대수 값은 -1.13이고, 후보 출력 토큰 in에 대응하는 확률 값의 대수 값은 -.098이고, EOS 토큰에 대응하는 확률 값의 대수 값은 -1.51이다.

[0121] 프로세서(200)는 복수의 제1 확률 값들(310)의 정렬을 통해 확률 값들의 순위를 획득할 수 있다. 프로세서(200)는 제1 후보 출력 토큰에 대응하는 제1 확률 값들이 미리 결정된 순위에 포함되는지 여부를 판단할 수 있다.

[0122] 미리 결정된 순위는 실험적으로 결정될 수 있다. 미리 결정된 순위는 실시예에 따라 달라질 수 있다. 미리 결정된 순위는 자연수일 수 있다. 도 3의 예시에서, 미리 결정된 순위는 2일 수 있다.

[0123] 프로세서(200)는 제1 확률 값이 미리 결정된 순위에 포함되지 않을 경우, 제1 확률 값을 감소시킬 수 있다. 예를 들어, 프로세서(200)는 제1 확률 값의 대수 값을 음의 무한대로 수정할 수 있다.

[0124] 도 3의 예시에서, 제1 후보 출력 토큰은 EOS 토큰일 수 있다. 도 3의 예시에서, EOS 토큰에 대응하는 확률 값(제1 확률 값)의 대수 값은 -1.51이다. 프로세서(200)는 제1 토큰인 EOS 토큰에 대응하는 확률 값이 미리 결정된 순위인 2위 내에 포함되지 않기 때문에, 제1 확률 값을 감소시킬 수 있다. 예를 들어, 프로세서(200)는 EOS 토큰에 대응하는 확률 값의 대수 값을 음의 무한대로 수정하여, 수정된 제1 확률 값(330)을 획득할 수 있다.

[0125] 프로세서(200)는 제1 확률 값들(310) 및 제2 확률 값들(350)에 기초하여 출력 토큰을 결정할 수 있다. 프로세서(200)는 제1 확률 값들(310) 및 제2 확률 값들(350)을 가중합할 수 있다. 프로세서(200)는 가중합한 값이 가장 큰 후보 출력 토큰을 출력 토큰으로 결정할 수 있다. 이 때, 제1 확률 값들(310)은 수정된 제1 확률 값

(330)이 반영된 확률 값을 의미할 수 있다.

[0126] 프로세서(200)는 상술한 바와 같이 복수의 제2 후보 출력 토큰에 대응하는 제2 확률 값들을 계산할 수 있다. 프로세서(200)는 수정된 제1 확률 값(330)이 반영된 제1 확률 값들과 제2 확률 값들(350)을 가중합할 수 있다.

[0127] EOS 토큰에 대응하는 확률 값의 대수 값이 음의 무한대가 되면 뉴럴 네트워크는 문장이 끝나지 않았다는 것으로 인식할 수 있다. 프로세서(200)는 EOS 토큰이 미리 결정된 순위에 포함되지 않을 경우, 제1 확률 값들(310) 중 EOS 토큰에 대응하는 확률 값을 음의 무한대로 수정함으로써, 가중합의 결과를 충분히 작은 값(즉, 충분히 큰 절대값을 갖는 음의 값)으로 수정할 수 있다.

[0128] 다른 실시예에서, 프로세서(200)는 제2 뉴럴 네트워크가 계산한 확률 값을 수정함으로써 가중합한 확률 값을 수정할 수 있다.

[0129] 프로세서(200)는 제1 후보 출력 토큰이 미리 결정된 순위 내에 포함되는 지를 판단하고, 제1 후보 출력 토큰에 대응하는 확률 값이 미리 결정된 순위 내에 포함되지 않는 경우 제1후보 출력 토큰에 대응하는 제2 확률 값(예를 들어, 제2 뉴럴 네트워크의 EOS 토큰에 대응하는 확률 값)을 수정할 수 있다. 예를 들어, 프로세서(200)는 제1 후보 출력 토큰에 대응하는 제2 확률 값의 대수 값을 음의 무한대로 수정할 수 있다.

[0130] 이를 통해, 프로세서(200)는 제1 뉴럴 네트워크를 통해 계산한 확률 값과 제2 뉴럴 네트워크를 통해 계산한 확률 값의 가중합을 수정함으로써 제1 후보 출력 토큰의 최종 출력 확률을 수정할 수 있다.

[0131] 또한, 프로세서(200)는 제1 후보 출력 토큰 자체를 제1 뉴럴 네트워크 또는 제2 뉴럴 네트워크에서 제외시킬 수도 있다. 이를 통해, 최종 출력에 제1 후보 출력 토큰이 포함되는 것을 방지하여 최종 출력이 문장 끝으로 인식되는 문제를 해결할 수 있다. 상술한 방식들을 통해, 프로세서(200)는 현재 입력 토큰에 의해 문장이 끝나는 것을 방지하여 추가적인 입력 토큰을 수신할 수 있다. 다시 말해, 프로세서(200)는 EOS 토큰에 대응하는 확률 값을 수정함으로써 실제 음성 신호의 끝이 아닌 지점을 문장의 끝이라고 인식하는 문제를 해결할 수 있다.

[0133] 이하에서, 도 4를 참조하여, 프로세서(200)가 비언어적 토큰을 포함하는 음성 신호를 처리하는 과정을 설명한다.

[0135] 도 4는 도 1에 도시된 음성 신호 처리 장치가 비언어적 토큰에 대응하는 확률을 생성하는 동작의 예를 나타낸다.

[0136] 도 4를 참조하면, 프로세서(200)는 언어 모델만으로는 추정할 수 없는 비언어적 심볼(symbol)이 포함된 음성 신호를 인식하기 위해서, 음성 인식 모델에서 비언어적 토큰의 등록 여부(예를 들어, 빔 등록 여부)를 판단하고, 등록되지 않았을 경우 비언어적 토큰 및 그에 대응하는 확률을 생성할 수 있다.

[0137] 이하에서, 예시적으로, 프로세서(200)가 비언어적 토큰 및 그에 대응하는 확률을 결정하는 과정을 설명한다.

[0138] 프로세서(200)는 입력 토큰에 기초하여 복수의 제1 후보 출력 토큰에 대응하는 제1 확률 값들(410)을 계산할 수 있다. 프로세서(200)는 입력 토큰에 기초하여 복수의 제2 후보 출력 토큰에 대응하는 제2 확률 값들(430)을 계산할 수 있다.

[0139] 프로세서(200)가 제1 확률 값들(410)들 및 제2 확률 값들(430)을 계산하는 과정은 도 3에서 설명한 것과 동일할 수 있다.

[0140] 프로세서(200)는 제1 확률 값들(410)들 및 제2 확률 값들(430) 중 적어도 하나에 기초하여 비언어적 토큰 및 비언어적 토큰에 대응하는 확률을 생성할 수 있다.

[0141] 프로세서(200)는 복수의 제1 후보 출력 토큰에 포함된 비언어적 토큰을 복수의 제2 후보 출력 토큰에 복사할 수 있다. 제1 후보 출력 토큰은 제1 뉴럴 네트워크에서 사용되는 후보 출력 토큰을 의미할 수 있다.

[0142] 후보 출력 토큰은 언어적 토큰 및 비언어적 토큰을 포함할 수 있다. 도 4의 예시에서, 복수의 제1 후보 출력 토큰은 from, at, in 및 기침(cough)을 포함할 수 있다. 이 중에서, from, at 및 in은 언어적 토큰이고, 기침은 비언어적 토큰일 수 있다.

[0143] 프로세서(200)는 제1 뉴럴 네트워크를 텍스트와 음성 신호가 결합된 데이터셋을 이용하여 학습시킴으로써 비언

어적 토큰에 대해서도 확률 값을 계산할 수 있다.

[0144] 반면, 제2 뉴럴 네트워크는 텍스트만을 이용하여 학습되어, 비언어적 음성 신호에 대응하는 토큰 및 그에 대응하는 확률 값을 가지지 않을 수 있다.

[0145] 프로세서(200)는 비언어적 토큰이 복수의 제2 후보 출력 토큰에 등록되어 있는지 여부를 판단할 수 있다. 프로세서(200)는 제2 후보 출력 토큰에 비언어적 토큰이 등록되어 있지 않은 경우, 이를 생성할 수 있다.

[0146] 프로세서(200)는 제1 뉴럴 네트워크의 비언어적 토큰 및 비언어적 토큰에 대응하는 확률 값에 기초하여 제2 뉴럴 네트워크에 비언어적 토큰 및 비언어적 토큰에 대응하는 확률 값을 생성할 수 있다.

[0147] 프로세서(200)는 복수의 제1 후보 출력 토큰에 포함된 비언어적 토큰을 복수의 제2 후보 출력 토큰에 복사할 수 있다. 프로세서(200)는 제1 확률 값들(410)에 포함된 비언어적 토큰에 대응하는 확률을 제2 확률 값들(430)들에 복사할 수 있다.

[0148] 도 4의 예시에서, 프로세서(200)는 복수의 제1 후보 출력 토큰 중에서 비언어적 토큰인 <기침> 토큰을 제2 후보 출력 토큰에 복사할 수 있다. 또한, 프로세서(200)는 제1 확률 값들(410) 중에서 <기침> 토큰에 대응하는 확률 값의 대수 값인 -0.26를 제2 확률 값들(430)에 복사할 수 있다.

[0149] 이에 따라, 프로세서(200)는 제2 뉴럴 네트워크에도 비언어적 토큰과 그에 대응하는 확률 값들을 생성할 수 있다.

[0150] 또는, 프로세서(200)는 복사된 후보 출력 토큰에 제2 확률 값들(430) 중 하나를 매핑할 수 있다. 예를 들어, 프로세서(200)는 제2 확률 값들(430) 중에서 가장 큰 값을 복사된 비언어적 토큰에 대응하는 확률로 매핑할 수 있다.

[0151] 프로세서(200)는 비언어적 토큰에 대응하는 확률에 기초하여 음성 신호를 처리할 수 있다. 프로세서(200)가 제1 확률 값들(410) 및 제2 확률 값들(430)을 이용하여 출력 토큰을 결정하는 과정은 도 3과 동일할 수 있다.

[0153] 도 5는 도 1에 도시된 음성 신호 처리 장치의 동작의 순서도를 나타낸다.

[0154] 도 5를 참조하면, 수신기(100)는 음성 신호에 기초한 입력 토큰을 수신할 수 있다(510).

[0155] 프로세서(200)는 제1 뉴럴 네트워크에서 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제1 확률 값들(310)을 계산할 수 있다(530). 제1 확률 값들(310)의 계산 과정은 도 2, 도 3에서 설명한 것과 동일할 수 있다.

[0156] 프로세서(200)는 제1 확률 값들(310)의 순위에 기초하여 제1 확률 값들(310) 중 적어도 하나의 값을 수정할 수 있다(550). 프로세서(200)는 복수의 후보 출력 토큰에 포함되는 제1 후보 출력 토큰에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지 여부를 판단할 수 있다.

[0157] 프로세서(200)는 판단 결과에 기초하여 제1 확률 값을 수정할 수 있다. 프로세서(200)는 제1 확률 값이 미리 결정된 순위에 포함되지 않을 경우, 제1 확률 값을 감소시킬 수 있다. 프로세서(200)는 제1 확률 값이 미리 결정된 순위 내에 포함될 경우, 제1 확률 값의 대수 값을 음의 무한대로 수정할 수 있다.

[0158] 프로세서(200)는 수정된 확률 값(330)에 기초하여 음성 신호를 처리할 수 있다(570). 프로세서(200)는 입력 토큰에 기초하여 복수의 후보 출력 토큰에 대응하는 제2 확률 값들을 계산할 수 있다.

[0159] 프로세서(200)는 계산한 제1 확률 값들 및 제2 확률 값들 중 적어도 하나에 기초하여 비언어적 토큰 및 비언어적 토큰에 대응하는 확률을 생성할 수 있다.

[0160] 프로세서(200)는 비언어적 토큰이 복수의 제2 후보 출력 토큰에 등록되어 있는지 여부를 판단할 수 있다. 프로세서(200)는 복수의 제1 후보 출력 토큰에 포함된 비언어적 토큰을 복수의 제2 후보 출력 토큰에 복사할 수 있다.

[0161] 프로세서(200)는 제1 확률 값들에 포함된 비언어적 토큰에 대응하는 확률을 제2 확률 값들에 복사할 수 있다. 또는, 프로세서(200)는 제2 확률 값들 중에서 가장 큰 값을 복사된 비언어적 토큰에 대응하는 확률로 매핑할 수 있다.

[0162] 프로세서(200)는 제1 확률 값들 및 제2 확률 값들에 기초하여 출력 토큰을 결정할 수 있다.

[0163] 프로세서(200)는 제1 확률 값들 및 제2 확률 값들 각각을 가중합할 수 있다. 프로세서(200)는 가중합한 값이 가장 큰 후보 출력 토큰을 출력 토큰으로 결정할 수 있다.

[0165] 도 6은 음성 신호 처리 방법의 예를 나타낸다.

[0166] 도 6을 참조하면, 프로세서(200)는 복수의 뉴럴 네트워크를 결합하여 음성 신호를 처리할 수 있다. 예를 들어, 프로세서(200)는 제1 뉴럴 네트워크 및 제2 뉴럴 네트워크를 이용하여 음성 신호를 처리할 수 있다.

[0167] 예를 들어, 제1 뉴럴 네트워크는 음성 인식 모델을 포함할 수 있고, 제2 뉴럴 네트워크는 언어 모델을 포함할 수 있다.

[0168] 프로세서(200)는 음성 인식 모델(예를 들어, 엔드-투-엔트(end-to-end) ASR(Automatic Speech Recognition)과 외부 언어 모델(external language model)의 확률을 결합할 때 언어 모델이 추정하기 어려운 토큰들을 고려해서 음성 신호를 처리하기 위해 언어 모델이 추정하기 어려운 토큰들에 대해 음성 인식 모델의 확률 값만으로 정렬하여 출력할 것인지 여부를 판단할 수 있다.

[0169] 프로세서(200)는 서치 알고리즘을 이용하여 음성 신호를 처리할 수 있다. 서치 알고리즘은 그리디 서치(greedy search) 알고리즘 및 빔 서치(beam search) 알고리즘을 포함할 수 있다.

[0170] 도 6의 예시는, 그리디 서치를 이용할 경우를 기준으로 설명하였으나, 실시예에 따라 프로세서(200)는 빔서치 알고리즘을 이용하여 음성 신호를 처리하는 것도 가능하다.

[0171] 빔 서치를 이용하는 경우, 프로세서(200)는 확률 값의 정렬을 통해서 후보 출력 토큰 들을 빔에 남겨야 하는지 여부를 판단할 수 있다. 프로세서(200)는 판단의 기준(criterion)에 따라 음성 인식 모델 또는 언어 모델의 후보 출력 토큰에 대응하는 확률을 조작하여 빔(beam)을 구성할 수 있다.

[0172] 다시 말해, 프로세서(200)는 제1 확률 값들(410)과 제2 확률 값들(430)의 가중합 전 단계에서 복수의 후보 출력 토큰 중에서 특정한 토큰의 확률 값이 미리 결정된 순위 내에 포함되는지를 판단할 수 있다. 미리 결정된 순위가 N일 경우, 미리 결정된 순위 내에 포함되는 경우를 N-best 내에 포함된다고 표현할 수 있다.

[0173] 프로세서(200)는 판단 결과에 따라 특정 토큰에 대응하는 확률을 수정하여 특정 토큰이 빔에 포함될지 여부를 결정할 수 있다.

[0174] 수신기(100)는 음성 신호에 기초한 입력 토큰을 수신할 수 있다(610). 프로세서(200)는 입력 토큰을 인코더(210)를 통해 인코딩할 수 있다(620).

[0175] 프로세서(200)는 i=0에 대응하는 시점에서, T에 <S>를 할당할 수 있다(630).

[0176] 프로세서(200)는 제1 뉴럴 네트워크의 디코더(230)를 이용하여 디코딩을 수행할 수 있다(640). 또한, 프로세서(200)는 제2 뉴럴 네트워크의 디코더를 이용하여 디코딩을 수행할 수 있다(650). 프로세서(200)는 제2 뉴럴 네트워크를 이용하여 제2 확률 값들을 계산할 수 있다.

[0177] 디코딩 과정은 제1 뉴럴 네트워크와 제2 뉴럴 네트워크에서 독립적으로 수행될 수 있다. 예를 들어, 음성 인식 모델과 외부 언어 모델이 독립적으로 후보 출력 토큰에 대응하는 확률 값들을 추정할 수 있다.

[0178] 프로세서(200)는 제1 뉴럴 네트워크를 이용하여 제1 확률 값들(310)을 계산할 수 있다(660). 프로세서(200)는 복수의 제1 후보 출력 토큰에 포함된 제1 후보 출력 토큰(예를 들어, EOS 토큰)에 대응하는 제1 확률 값이 미리 결정된 순위에 포함되는지(즉, N-best 내에 들어오는지) 판단할 수 있다(670).

[0179] 프로세서(200)는 제1 확률 값이 미리 결정된 순위에 포함되지 않는 경우, 제1 확률 값의 대수 값을 음의 무한대로 수정할 수 있다(680).

[0180] 이를 통해, 프로세서(200)는 제1 후보 출력 토큰이 빔으로 선택되는 것을 방지함으로써, 긴 문장을 인식할 때, 중간에서 인식이 끊어지는 문제를 해결할 수 있다.

[0181] 프로세서(200)는 수정된 제1 확률 값(330)을 포함하는 제1 확률 값들(310)과 제2 확률 값들(350)의 가중합을 계산할 수 있다(690). 여기서, 프로세서(200)는 미리 결정된 하이퍼-파라미터(hyper-parameter)를 이용하여 가중합을 수행할 수 있다.

[0182] 프로세서(200)는 EOS 토큰에 대응하는 확률 값이 최대 확률 값인지를 판단할 수 있다(700). EOS 토큰에 대응하

는 확률 값이 최대인 경우 추정을 종료하고, 최대 값이 아닐 경우, 다음 시점에 대한 입력 토큰을 처리할 수 있다(710). 다시 말해, i=i+1 및 T=T+[w_i]를 수행할 수 있다.

[0183] 다음 시점에 대하여 출력 토큰을 결정할 때, 640, 650 단계를 반복적으로 수행할 수 있다.

[0185] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

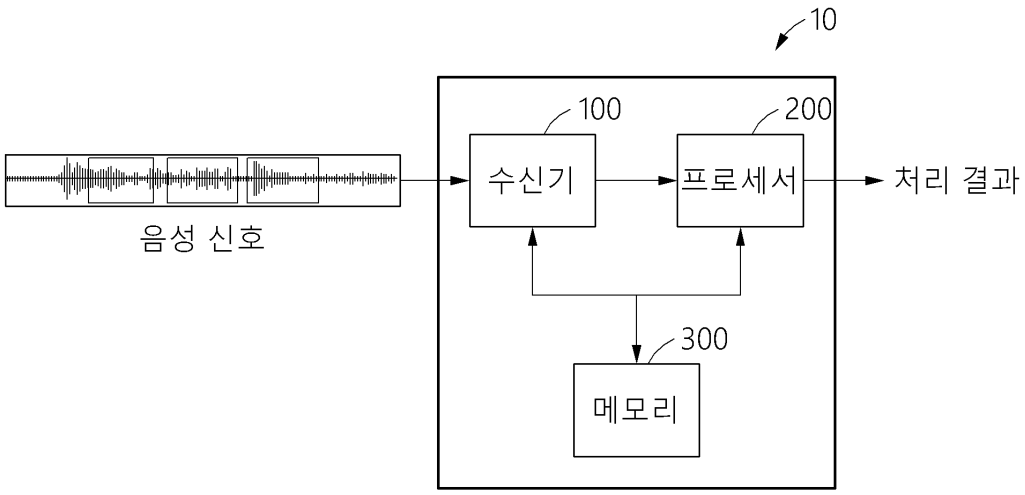
[0187] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0189] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

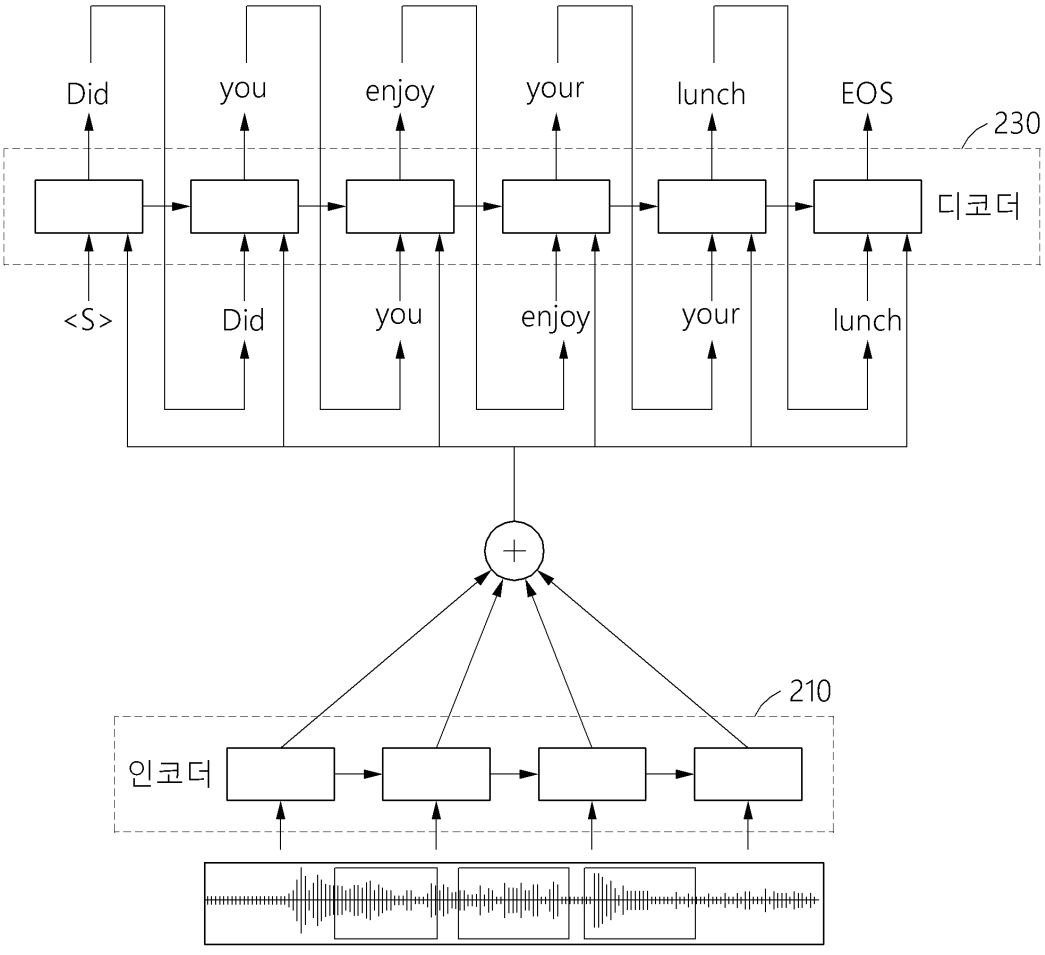
[0190] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

도면

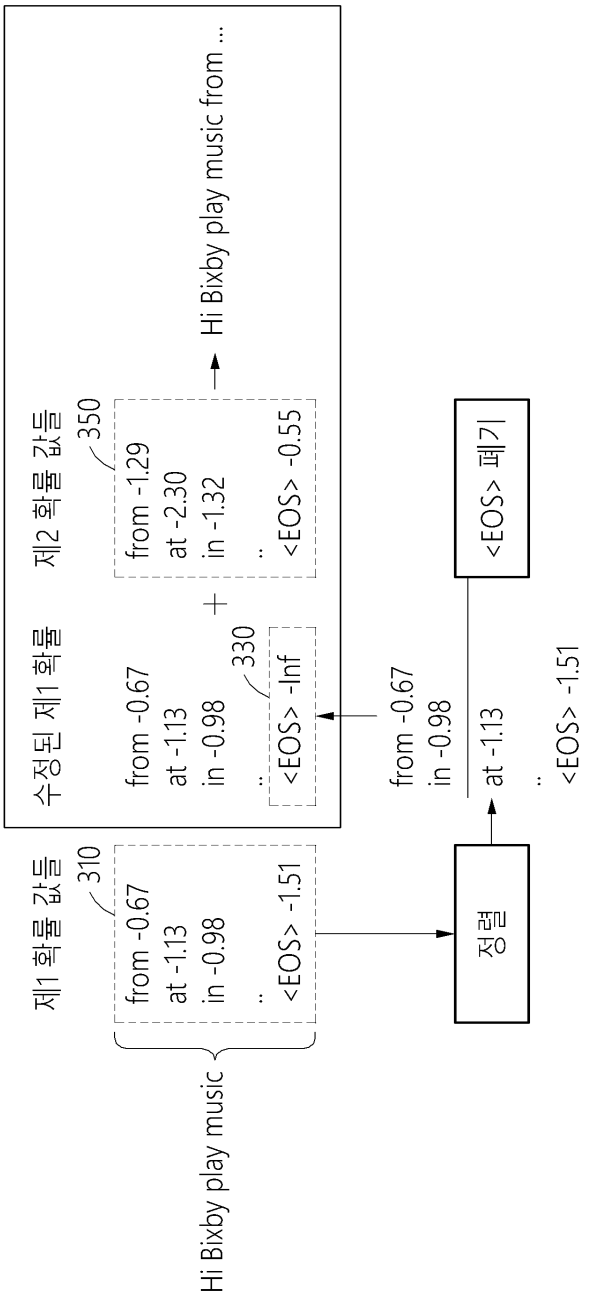
도면1


10
100
200
수신기
프로세서
처리 결과
음성 신호
300
메모리

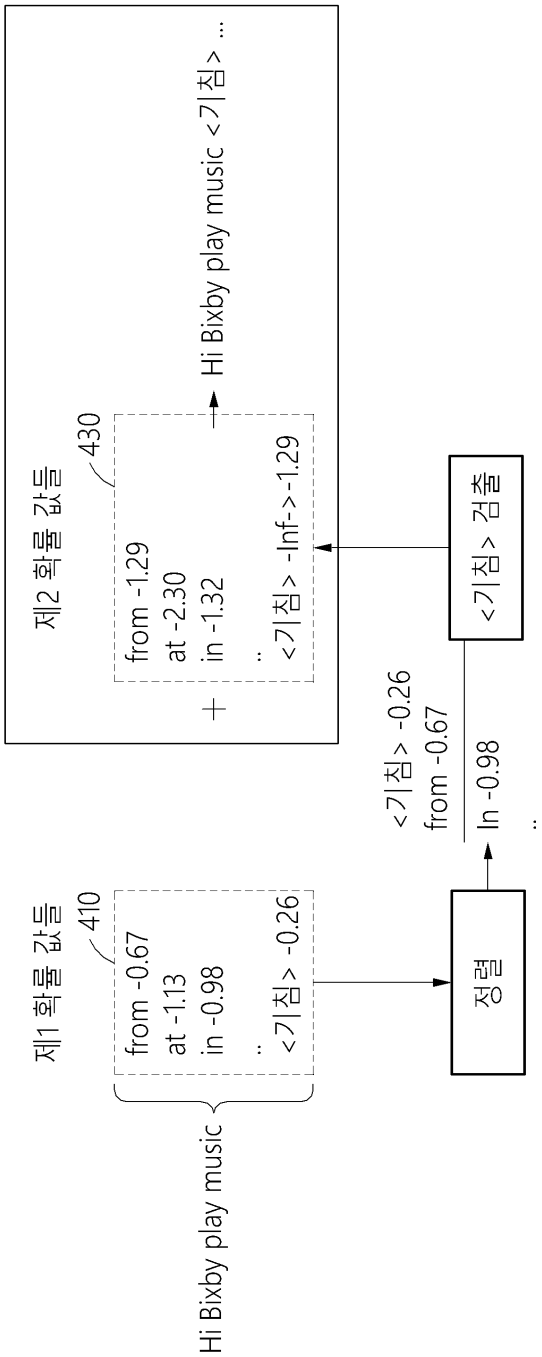
도면2


Did
you
enjoy
your
lunch
EOS
230
디코더
<S>
Did
you
enjoy
your
lunch
210
인코더

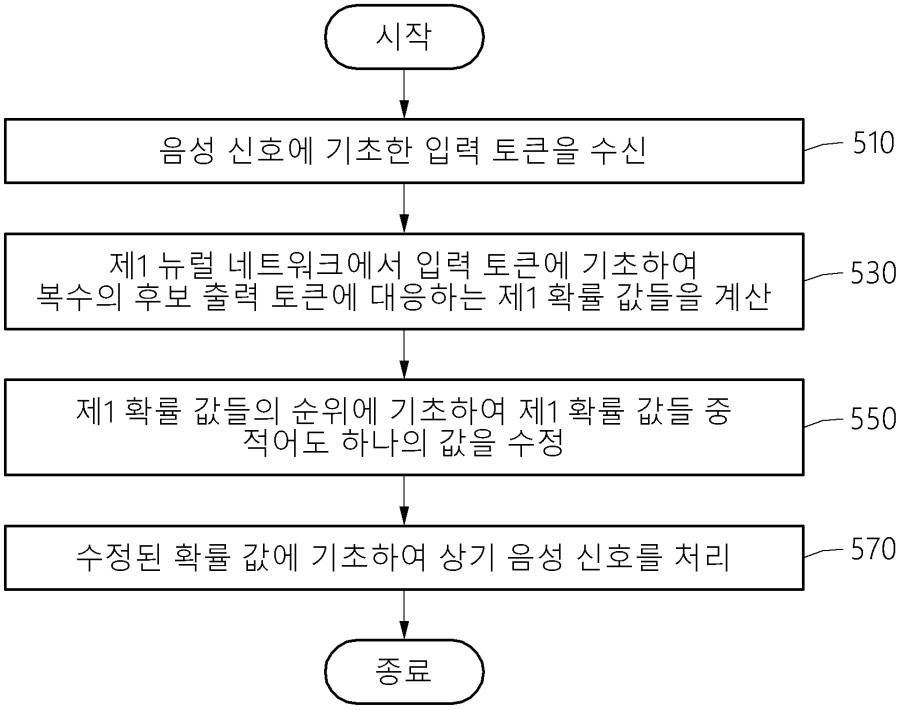
도면3


제1 확률 값들
310
from -0.67
at -1.13
in -0.98
..
<EOS> -1.51
Hi Bixby play music
수정된 제1 확률
from -0.67
at -1.13
in -0.98
..
330
<EOS> -Inf
제2 확률 값들
350
from -1.29
at -2.30
in -1.32
..
<EOS> -0.55
Hi Bixby play music from ...
정렬
from -0.67
in -0.98
at -1.13
..
<EOS> -1.51
<EOS> 떼기

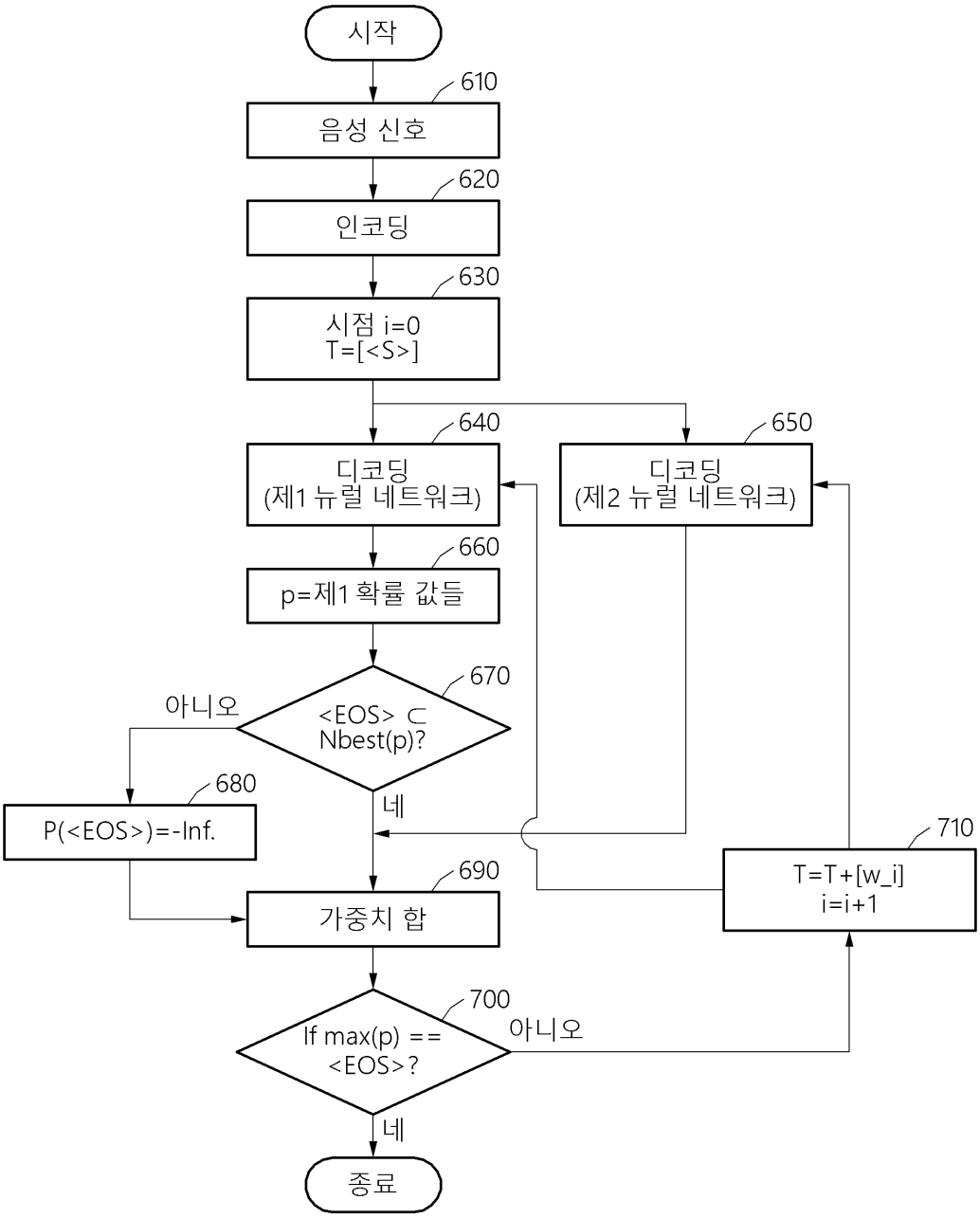
도면4


제2 확률 값들
430
from -1.29
at -2.30
in -1.32
..
<기침> -Inf->-1.29
Hi Bixby play music <기침> ...
제1 확률 값들
410
from -0.67
at -1.13
in -0.98
..
<기침> -0.26
Hi Bixby play music
정렬
<기침> -0.26
from -0.67
In -0.98
..
<기침> 검출

도면5


시작
음성 신호에 기초한 입력 토큰을 수신
510
제1 뉴럴 네트워크에서 입력 토큰에 기초하여
복수의 후보 출력 토큰에 대응하는 제1 확률 값들을 계산
530
제1 확률 값들의 순위에 기초하여 제1 확률 값들 중
적어도 하나의 값을 수정
550
수정된 확률 값에 기초하여 상기 음성 신호를 처리
570
종료

도면6


시작
610
음성 신호
620
인코딩
630
시점 i=0
T=[<S>]
640
디코딩
(제1 뉴럴 네트워크)
650
디코딩
(제2 뉴럴 네트워크)
660
p=제1 확률 값들
670
<EOS> ⊂
Nbest(p)?
아니오
네
680
P(<EOS>)=-Inf.
690
가중치 합
710
T=T+[w_i]
i=i+1
700
If max(p) ==
<EOS>?
아니오
네
종료