(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년06월20일
(11) 등록번호 10-2410820
(24) 등록일자 2022년06월15일

(51) 국제특허분류(Int. Cl.)
G10L 15/16 (2006.01) *G06N 3/08* (2006.01)
G10L 15/06 (2006.01)
(52) CPC특허분류
G10L 15/16 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2017-0103044
(22) 출원일자 2017년08월14일
심사청구일자 2020년08월13일
(65) 공개번호 10-2019-0018278
(43) 공개일자 2019년02월22일
(56) 선행기술조사문헌
KR1020170046751 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
유상현
서울특별시 광진구 구의강변로 94, 603동 1803호 (구의3동, 현대6차아파트)
(74) 대리인
특허법인 무한

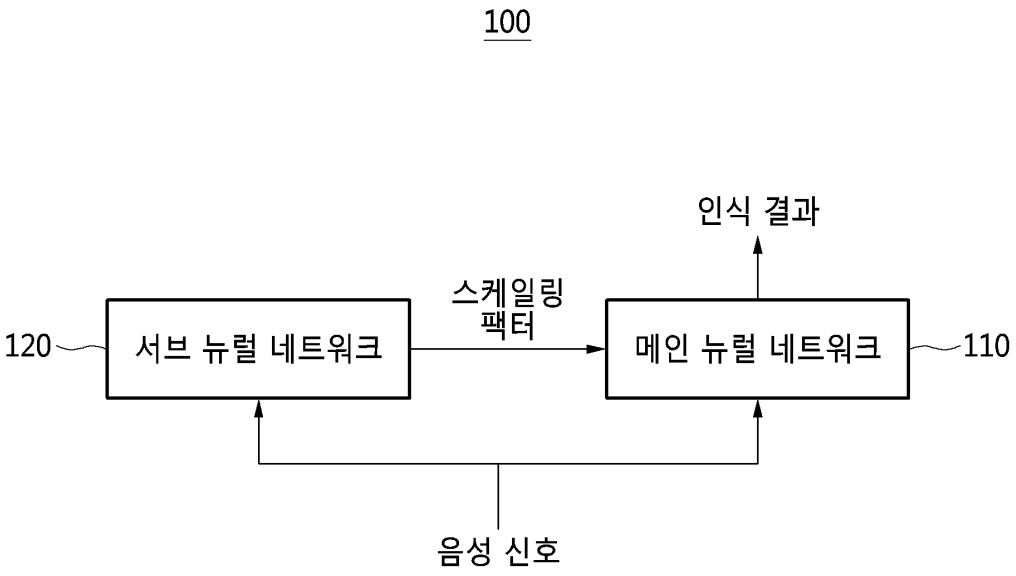
전체 청구항 수 : 총 18 항 심사관 : 이현중

(54) 발명의 명칭 **뉴럴 네트워크를 이용한 인식 방법 및 장치 및 상기 뉴럴 네트워크를 트레이닝하는 방법 및 장치**

(57) 요 약

룩어헤드 컨볼루션 레이어를 포함한 뉴럴 네트워크에 기초한 인식 방법 및 장치, 및 상기 뉴럴 네트워크를 트레이닝하는 방법 및 장치가 개시된다. 개시된 인식 방법은 음성 신호가 입력된 서브 뉴럴 네트워크로부터 스케일링 팩터를 획득하고, 스케일링 팩터에 기초하여 메인 뉴럴 네트워크에서 고려할 미래 컨텍스트 길이를 결정하며, 결정된 미래 컨텍스트 길이가 적용된 메인 뉴럴 네트워크에 음성 신호가 입력됨에 따라 메인 뉴럴 네트워크에서 출력되는 음성 신호의 인식 결과를 획득한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 15/063 (2013.01)

명 세 서

청구범위

청구항 1

음성 신호가 입력된 서브 뉴럴 네트워크(sub neural network)로부터 스케일링 팩터(scaling factor)를 획득하는 단계;

상기 스케일링 팩터에 기초하여 상기 음성 신호의 명확성이 낮을수록 메인 뉴럴 네트워크(main neural network)에서 고려할 미래 컨텍스트 길이(future context size)를 크게 결정하는 단계; 및

상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 음성 신호가 입력됨에 따라 상기 메인 뉴럴 네트워크에서 출력되는 상기 음성 신호의 인식 결과를 획득하는 단계

를 포함하는 인식 방법.

청구항 2

제1항에 있어서,

상기 스케일링 팩터를 획득하는 단계는

상기 서브 뉴럴 네트워크를 이용하여, 상기 음성 신호가 인식될 때 상기 메인 뉴럴 네트워크에서 고려될 미래 컨텍스트의 시점별 중요도를 나타내는 스케일링 팩터를 획득하는, 인식 방법.

청구항 3

제2항에 있어서,

상기 스케일링 팩터를 획득하는 단계는

상기 음성 신호에 포함된 노이즈 정도 및 상기 음성 신호에 포함된 사용자 발음의 정확도 중 적어도 하나에 기초하여 결정된 스케일링 팩터를 획득하는, 인식 방법.

청구항 4

제2항에 있어서,

상기 스케일링 팩터는

상기 음성 신호의 명확성이 낮을수록 미리 결정된 임계치보다 큰 값을 가지는 성분이 많아지도록 결정되는, 인식 방법.

청구항 5

삭제

청구항 6

제1항에 있어서,

상기 미래 컨텍스트 길이를 결정하는 단계는

상기 스케일링 팩터에 포함된 성분들의 값과 미리 결정된 임계치 간의 비교를 통해 상기 미래 컨텍스트 길이를 결정하는, 인식 방법.

청구항 7

제6항에 있어서,

상기 미래 컨텍스트 길이를 결정하는 단계는

상기 미리 결정된 임계치보다 큰 값을 가지는 상기 스케일링 팩터의 성분들 중 가장 높은 차원에 기초하여 상기 미래 컨텍스트 길이를 결정하는, 인식 방법.

청구항 8

제1항에 있어서,

상기 음성 신호의 인식 결과를 획득하는 단계는

상기 메인 뉴럴 네트워크에 포함된 룩어헤드 컨볼루션 레이어(lookahead convolution layer)의 미래 컨텍스트 길이를 상기 결정된 미래 컨텍스트 길이로 조절하는 단계;

상기 조절된 룩어헤드 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크에 상기 음성 신호를 입력하는 단계; 및

상기 조절된 룩어헤드 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크로부터 출력되는 음성 신호의 인식 결과를 획득하는 단계

를 포함하는 인식 방법.

청구항 9

제1항에 있어서,

상기 미래 컨텍스트 길이를 결정하는 단계는

상기 음성 신호에 대한 첫 번째 윈도우에서 획득한 스케일링 팩터에 기초하여 상기 미래 컨텍스트 길이를 결정하고,

상기 음성 신호의 인식 결과를 획득하는 단계는

상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크를 이용하여, 상기 음성 신호에 적용된 복수의 윈도우들에 대한 인식 결과를 획득하는, 인식 방법.

청구항 10

제1항에 있어서,

상기 메인 뉴럴 네트워크는 룩어헤드 컨볼루션 레이어를 포함한 단방향 리커런트 뉴럴 네트워크(Unidirectional Recurrent Neural Network; Unidirectional RNN)인, 인식 방법.

청구항 11

제1항에 있어서,

상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크는 함께 트레이닝되는, 인식 방법.

청구항 12

트레이닝 입력이 입력된 서브 뉴럴 네트워크로부터 스케일링 팩터를 획득하는 단계;

상기 스케일링 팩터에 기초하여 상기 트레이닝 입력의 명확성이 낮을수록 메인 뉴럴 네트워크에서 고려할 미래 컨텍스트 길이를 크게 결정하는 단계; 및

상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 트레이닝 입력이 입력됨에 따라 상기 메인 뉴럴 네트워크에서 상기 트레이닝 입력에 매핑된 트레이닝 출력이 출력되도록, 상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크를 트레이닝시키는 단계

를 포함하는 트레이닝 방법.

청구항 13

제12항에 있어서,

상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크를 트레이닝시키는 단계는

상기 메인 뉴럴 네트워크와 상기 서브 뉴럴 네트워크를 동시에 트레이닝시키는, 트레이닝 방법.

청구항 14

제1항 내지 제4항 및 제6항 내지 제13항 중 어느 한 항의 방법을 수행하기 위한 명령어들을 포함하는 하나 이상의 프로그램을 저장한 컴퓨터 판독 가능 저장매체.

청구항 15

프로세서; 및

상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리

를 포함하고,

상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는

음성 신호가 입력된 서브 뉴럴 네트워크로부터 스케일링 팩터를 획득하고, 상기 스케일링 팩터에 기초하여 상기 음성 신호의 명확성이 낮을수록 메인 뉴럴 네트워크에서 고려할 미래 컨텍스트 길이를 크게 결정하며, 상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 음성 신호가 입력됨에 따라 상기 메인 뉴럴 네트워크에서 출력되는 상기 음성 신호의 인식 결과를 획득하는, 인식 장치.

청구항 16

제15항에 있어서,

상기 프로세서는

상기 서브 뉴럴 네트워크를 이용하여, 상기 음성 신호가 인식될 때 상기 메인 뉴럴 네트워크에서 고려될 미래 컨텍스트의 시점별 중요도를 나타내는 스케일링 팩터를 획득하는, 인식 장치.

청구항 17

제16항에 있어서,

상기 프로세서는

상기 음성 신호에 포함된 노이즈 정도 및 상기 음성 신호에 포함된 사용자 발음의 정확도 중 적어도 하나에 기초하여 결정된 스케일링 팩터를 획득하는, 인식 장치.

청구항 18

삭제

청구항 19

제15항에 있어서,

상기 프로세서는

상기 스케일링 팩터에 포함된 성분들의 값과 미리 결정된 임계치 간의 비교를 통해 상기 미래 컨텍스트 길이를 결정하는, 인식 장치.

청구항 20

제15항에 있어서,

상기 프로세서는

상기 메인 뉴럴 네트워크에 포함된 룩어헤드 컨볼루션 레이어의 미래 컨텍스트 길이를 상기 결정된 미래 컨텍스트 길이로 조절하고, 상기 조절된 룩어헤드 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크에 상기 음성 신호를 입력하며, 상기 조절된 룩어헤드 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크로부터 출력된 음성 신호의 인식 결과를 획득하는, 인식 장치.

발명의 설명

기 술 분 야

[0001] 아래 실시예들은 룩어헤드 컨볼루션 레이어를 포함한 뉴럴 네트워크에 기초한 인식 방법 및 장치, 및 상기 뉴럴 네트워크를 트레이닝하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 들어, 입력 패턴을 특정 그룹으로 분류하는 문제를 해결하는 방안으로써, 인간이 지니고 있는 효율적인 패턴 인식 방법을 실제 컴퓨터에 적용시키려는 연구가 활발히 진행되고 있다. 이러한 연구 중 하나로, 인간의 생물학적 신경 세포의 특성을 수학적 표현에 의해 모델링한 인공 뉴럴 네트워크(artificial neural network)에 대한 연구가 있다. 입력 패턴을 특정 그룹으로 분류하는 문제를 해결하기 위해, 인공 뉴럴 네트워크는 인간이 가지고 있는 학습이라는 능력을 모방한 알고리즘을 이용한다. 이 알고리즘을 통하여 인공 뉴럴 네트워크는 입력 패턴과 출력 패턴들 사이의 사상(mapping)을 생성해낼 수 있는데, 이를 인공 뉴럴 네트워크가 학습 능력이 있다고 표현한다. 또한, 인공 뉴럴 네트워크는 학습된 결과에 기초하여 학습에 이용되지 않았던 입력 패턴에 대하여 비교적 올바른 출력을 생성할 수 있는 일반화 능력을 가지고 있다.

발명의 내용

과제의 해결 수단

[0003] 일실시예에 따른 인식 방법은 음성 신호가 입력된 서브 뉴럴 네트워크(sub neural network)로부터 스케일링 팩터(scaling factor)를 획득하는 단계; 상기 스케일링 팩터에 기초하여 메인 뉴럴 네트워크(main neural network)에서 고려할 미래 컨텍스트 길이(future context size)를 결정하는 단계; 및 상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 음성 신호가 입력됨에 따라 상기 메인 뉴럴 네트워크에서 출력되는 상기 음성 신호의 인식 결과를 획득하는 단계를 포함한다.

[0004] 일실시예에 따른 인식 방법에서 상기 스케일링 팩터를 획득하는 단계는 상기 서브 뉴럴 네트워크를 이용하여, 상기 음성 신호가 인식될 때 상기 메인 뉴럴 네트워크에서 고려될 미래 컨텍스트의 시점별 중요도를 나타내는 스케일링 팩터를 획득할 수 있다.

[0005] 일실시예에 따른 인식 방법에서 상기 스케일링 팩터를 획득하는 단계는 상기 음성 신호에 포함된 노이즈 정도 및 상기 음성 신호에 포함된 사용자 발음의 정확도 중 적어도 하나에 기초하여 결정된 스케일링 팩터를 획득할 수 있다.

[0006] 일실시예에 따른 인식 방법에서 상기 스케일링 팩터는 상기 음성 신호의 명확성이 낮을수록 미리 결정된 임계치보다 큰 값을 가지는 성분이 많아지도록 결정될 수 있다.

[0007] 일실시예에 따른 인식 방법에서 상기 미래 컨텍스트 길이를 결정하는 단계는 상기 스케일링 팩터에 기초하여 상기 음성 신호의 명확성이 낮을수록 상기 메인 뉴럴 네트워크에서 고려할 상기 미래 컨텍스트 길이를 크게 결정할 수 있다.

[0008] 일실시예에 따른 인식 방법에서 상기 미래 컨텍스트 길이를 결정하는 단계는 상기 스케일링 팩터에 포함된 성분들의 값과 미리 결정된 임계치 간의 비교를 통해 상기 미래 컨텍스트 길이를 결정할 수 있다.

[0009] 일실시예에 따른 인식 방법에서 상기 미래 컨텍스트 길이를 결정하는 단계는 상기 미리 결정된 임계치보다 큰 값을 가지는 상기 스케일링 팩터의 성분들 중 가장 높은 차원에 기초하여 상기 미래 컨텍스트 길이를 결정할 수 있다.

[0010] 일실시예에 따른 인식 방법에서 상기 음성 신호의 인식 결과를 획득하는 단계는 상기 메인 뉴럴 네트워크에 포함된 룩어헤드 컨볼루션 레이어(lookahead convolution layer)의 미래 컨텍스트 길이를 상기 결정된 미래 컨텍스트 길이로 조절하는 단계; 상기 조절된 룩어헤드 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크에 상기 음성 신호를 입력하는 단계; 및 상기 조절된 룩어헤드 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크로부터 출력되는 음성 신호의 인식 결과를 획득하는 단계를 포함할 수 있다.

[0011] 일실시예에 따른 인식 방법에서 상기 미래 컨텍스트 길이를 결정하는 단계는 상기 음성 신호에 대한 첫 번째 윈도우에서 획득한 스케일링 팩터에 기초하여 상기 미래 컨텍스트 길이를 결정하고, 상기 음성 신호의 인식 결과

를 획득하는 단계는 상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크를 이용하여, 상기 음성 신호에 적용된 복수의 윈도우들에 대한 인식 결과를 획득할 수 있다.

[0012] 일실시예에 따른 인식 방법에서 상기 메인 뉴럴 네트워크는 룩어헤드 컨볼루션 레이어를 포함한 단방향 리커런트 뉴럴 네트워크(Unidirectional Recurrent Neural Network; Unidirectional RNN)일 수 있다.

[0013] 일실시예에 따른 인식 방법에서 상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크는 함께 트레이닝될 수 있다.

[0014] 일실시예에 따른 트레이닝 방법은 트레이닝 입력이 입력된 서브 뉴럴 네트워크로부터 스케일링 팩터를 획득하는 단계; 상기 스케일링 팩터에 기초하여 메인 뉴럴 네트워크에서 고려할 미래 컨텍스트 길이를 결정하는 단계; 및 상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 트레이닝 입력이 입력됨에 따라 상기 메인 뉴럴 네트워크에서 상기 트레이닝 입력에 매핑된 트레이닝 출력이 출력되도록, 상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크를 트레이닝시키는 단계를 포함한다.

[0015] 일실시예에 따른 트레이닝 방법에서 상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크를 트레이닝시키는 단계는 상기 메인 뉴럴 네트워크와 상기 서브 뉴럴 네트워크를 동시에 트레이닝시킬 수 있다.

[0016] 일실시예에 따른 인식 장치는 프로세서; 및 상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리를 포함하고, 상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는 음성 신호가 입력된 서브 뉴럴 네트워크로부터 스케일링 팩터를 획득하고, 상기 스케일링 팩터에 기초하여 메인 뉴럴 네트워크에서 고려할 미래 컨텍스트 길이를 결정하며, 상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 음성 신호가 입력됨에 따라 상기 메인 뉴럴 네트워크에서 출력되는 상기 음성 신호의 인식 결과를 획득한다.

[0017] 일실시예에 따른 트레이닝 장치는 프로세서; 및 상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리를 포함하고, 상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는 트레이닝 입력이 입력된 서브 뉴럴 네트워크로부터 스케일링 팩터를 획득하고, 상기 스케일링 팩터에 기초하여 메인 뉴럴 네트워크에서 고려할 미래 컨텍스트 길이를 결정하며, 상기 결정된 미래 컨텍스트 길이가 적용된 상기 메인 뉴럴 네트워크에 상기 트레이닝 입력이 입력됨에 따라 상기 메인 뉴럴 네트워크에서 상기 트레이닝 입력에 매핑된 트레이닝 출력이 출력되도록, 상기 메인 뉴럴 네트워크 및 상기 서브 뉴럴 네트워크를 트레이닝시킨다.

도면의 간단한 설명

[0018] 도 1은 일실시예에 따른 인식 장치에서 음성 신호가 인식되는 과정을 나타낸 도면이다.

도 2는 일실시예에 따라 메인 뉴럴 네트워크를 나타낸 도면이다.

도 3은 일실시예에 따라 스케일링 팩터가 메인 뉴럴 네트워크에 적용되는 과정을 나타낸 도면이다.

도 4는 일실시예에 따라 서브 뉴럴 네트워크를 나타낸 도면이다.

도 5는 일실시예에 따라 스케일링 팩터에 기초하여 미래 컨텍스트 길이가 결정되는 과정을 나타낸 도면이다.

도 6은 일실시예에 따라 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크를 트레이닝하는 과정을 나타낸 도면이다.

도 7은 일실시예에 따라 스케일링 팩터를 설명하기 위한 도면이다.

도 8은 일실시예에 따라 스케일링 팩터를 결정하는 과정을 나타낸 도면이다.

도 9는 일실시예에 따른 인식 방법을 나타낸 도면이다.

도 10은 일실시예에 따른 트레이닝 방법을 나타낸 도면이다.

도 11은 일실시예에 따른 인식 장치를 나타낸 도면이다.

도 12는 일실시예에 따른 트레이닝 장치를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0019] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 실시될 수 있다. 따라서, 실시예들은 특정한 개시형태로 한정되는 것이 아니며, 본 명세서의 범

위는 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0020] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0021] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0022] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0023] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0024] 하기에서 설명될 실시예들은 음성을 인식하거나 음성 인식을 위한 뉴럴 네트워크를 트레이닝시키는 데 사용될 수 있다. 실시예들은 퍼스널 컴퓨터, 랩톱 컴퓨터, 태블릿 컴퓨터, 스마트 폰, 스마트 가전 기기, 지능형 자동차, 키오스크, 웨어러블 장치 등 다양한 형태의 제품으로 구현될 수 있다. 예를 들어, 실시예들은 스마트 폰, 모바일 기기, 스마트 홈 시스템 등에서 사용자의 음성을 인식하거나 해당 장치에서 사용자의 음성을 인식하기 위한 뉴럴 네트워크를 트레이닝시키는 데 적용될 수 있다. 또한, 실시예들은 음성 인식을 통한 장치 제어에도 적용될 수 있다. 이하, 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0025] 도 1은 일실시예에 따른 인식 장치에서 음성 신호가 인식되는 과정을 나타낸 도면이다.

[0026] 도 1을 참조하면, 일실시예에 따른 인식 장치(100)는 메인 뉴럴 네트워크(110) 및 서브 뉴럴 네트워크(120)를 포함한다.

[0027] 일실시예에 따른 인식 장치(100)는 메인 뉴럴 네트워크(110) 및 서브 뉴럴 네트워크(120)를 이용하여 음성 신호를 인식하는 장치로서, 아래에서 설명되는 적어도 하나의 동작을 위한 명령어들을 저장하는 메모리 및 해당 명령어들을 실행하는 프로세서를 포함할 수 있다. 메인 뉴럴 네트워크(110) 및 서브 뉴럴 네트워크(120)는 인식 장치(100)의 메모리에 명령어 형태로 저장될 수 있다.

[0028] 일실시예에 따른 메인 뉴럴 네트워크(110) 및 서브 뉴럴 네트워크(120)는 복수의 레이어들을 포함하는 뉴럴 네트워크일 수 있다. 복수의 레이어들 각각은 복수의 뉴런들을 포함할 수 있다. 이웃한 레이어들의 뉴런들은 시냅스들로 연결될 수 있다. 학습에 따라 시냅스들에는 가중치들이 부여될 수 있고, 파라미터들은 이러한 가중치들을 포함할 수 있다.

[0029] 뉴럴 네트워크의 복수의 레이어들에는 입력 레이어, 히든 레이어 및 출력 레이어가 포함될 수 있다. 예를 들어, 입력 레이어는 트레이닝 또는 인식의 대상이 되는 입력을 수신하여 히든 레이어로 전달할 수 있고, 출력 레이어는 히든 레이어의 뉴런들로부터 수신한 신호에 기초하여 뉴럴 네트워크의 출력을 생성할 수 있다. 히든 레이어는 입력 레이어와 출력 레이어 사이에 위치한 중간 레이어로서, 입력 레이어를 통해 전달된 학습 데이터 또는 인식의 대상이 되는 데이터를 예측하기 쉬운 값으로 변화시킬 수 있다.

[0030] 일실시예에 따른 메인 뉴럴 네트워크(110)는 입력된 음성 신호에 대응하는 인식 결과를 스케일링 팩터(scaling factor)에 기초하여 출력할 수 있다. 음성 신호는 사용자 음성을 포함한 신호로서 음성 시퀀스로 지칭될 수도 있으며, 복수의 음성 프레임들로 구분될 수 있다. 여기서, 프레임은 윈도우로도 지칭될 수 있다.

[0031] 메인 뉴럴 네트워크(110)는 음향 모델(acoustic model)를 포함할 수 있다. 음향 모델은 입력되는 음성 신호의 각 프레임이 어떤 음소에 가까운지에 대한 정보를 출력할 수 있다. 음향 모델에서 출력되는 정보를 이용하여, 입력된 음성 신호가 어떤 단어에 가까운지 또는 어떤 문장에 가까운지를 알 수 있다.

[0032] 메인 뉴럴 네트워크(110)는 룩어헤드 컨볼루션 레이어(lookahead convolution layer)를 포함할 수 있다. 룩어

헤드 컨볼루션 레이어는 단방향 리커런트 뉴럴 네트워크(Unidirectional Recurrent Neural Network; Unidirectional RNN)에서 일정 길이(size)의 미래 컨텍스트(future context)를 더 이용하여 출력을 생성하는 레이어를 나타낼 수 있다. 룩어헤드 컨볼루션 레이어에 대해서는 도 2를 참조하여 후술한다.

[0033] 메인 뉴럴 네트워크(110)에서 고려되는 미래 컨텍스트 길이(future context size)는 서브 뉴럴 네트워크(120)로부터 수신한 스케일링 팩터에 기초하여 결정될 수 있다.

[0034] 일실시예에 따른 서브 뉴럴 네트워크(120)는 입력되는 음성 신호에 기초하여 스케일링 팩터를 결정할 수 있다. 서브 뉴럴 네트워크(120)는 리커런트 뉴럴 네트워크(RNN), 컨볼루셔널 뉴럴 네트워크(Convolutional Neural Network) 또는 딥 뉴럴 네트워크(Deep Neural Network; DNN)을 포함할 수 있다.

[0035] 일실시예에 따른 스케일링 팩터는 음성 신호가 인식될 때 메인 뉴럴 네트워크에서 고려될 미래 컨텍스트의 시점별 중요도를 나타낼 수 있다. 스케일링 팩터는 음성 신호의 명확성(clarity)에 기초하여 결정될 수 있다. 예컨대, 스케일링 팩터는 음성 신호에 포함된 노이즈 정도 및 음성 신호에 포함된 사용자 발음의 정확도 중 적어도 하나에 기초하여 결정될 수 있다. 이러한 스케일링 팩터에 기초하여 메인 뉴럴 네트워크(110)에서 고려되는 미래 컨텍스트 길이가 결정될 수 있다.

[0036] 예를 들어, 음성 신호에 포함된 노이즈가 많거나, 음성 신호에 포함된 사용자 발음이 부정확한 경우, 해당 음성 신호를 정확하게 인식하기 위해서는 보다 많은 미래 컨텍스트가 고려될 필요가 있을 수 있다. 반면, 음성 신호에 포함된 노이즈가 적거나, 음성 신호에 포함된 사용자 발음이 정확한 경우, 적은 미래 컨텍스트만으로도 해당 음성 신호를 정확하게 인식할 수 있다.

[0037] 이와 같이, 스케일링 팩터를 이용하여 메인 뉴럴 네트워크(110)에서 고려할 미래 컨텍스트 길이를 적응적으로 조절함으로써, 음성 인식을 효율적으로 수행할 수 있다. 또한, 스케일링 팩터에 기초하여 최소한의 미래 컨텍스트만을 고려하여 음성 인식을 수행함으로써, 높은 정확도로 음성 인식을 수행하면서도 반응속도를 최대화할 수 있다.

[0038] 도 2는 일실시예에 따라 메인 뉴럴 네트워크를 나타낸 도면이다.

[0039] 도 2를 참조하면, 일실시예에 따른 메인 뉴럴 네트워크(110)는 리커런트 레이어(Recurrent Layer)(111) 및 룩어헤드 컨볼루션 레이어(113)를 포함할 수 있다. 도 2에는 설명의 편의를 위해 메인 뉴럴 네트워크(110) 내에 두 개의 히든 레이어들이 도시되어 있으나, 메인 뉴럴 네트워크(110)는 하나 이상의 히든 레이어를 제한 없이 포함할 수 있다. 또한, 히든 레이어는 하나 이상의 히든 노드를 제한 없이 포함할 수 있다.

[0040] 메인 뉴럴 네트워크(110)에서 인접한 레이어에 속한 노드들은 시냅스를 통해 서로 연결될 수 있고, 시냅스에는 미리 트레이닝된 가중치가 부여될 수 있다.

[0041] 일실시예에 따른 리커런트 레이어(111)는 회기 루프를 가질 수 있으며, 도 2에서는 설명의 편의를 위해 리커런트 레이어(111)가 펼쳐진(unfolded) 상태로 도시될 수 있다. 예를 들어, 시점 t의 리커런트 레이어(111)의 출력 x_t은 시점 t+1의 리커런트 레이어(111)에 다시 입력됨으로써, 시점 t+1에서 새로운 출력 x_{t+1}이 출력될 수 있다.

[0042] 일실시예에 따른 룩어헤드 컨볼루션 레이어(113)는 미리 결정된 길이의 미래 컨텍스트를 고려할 수 있다. 도 2에서는 설명의 편의를 위해 미래 컨텍스트 길이 τ가 2인 예시가 도시된다.

[0043] 다시 말해, 시점 t의 룩어헤드 컨볼루션 레이어(113)는 시점 t의 리커런트 레이어(111)의 출력 x_t뿐만 아니라 시점 t+1, t+2의 리커런트 레이어(111)의 출력 x_{t+1}, x_{t+2}을 더 고려하여 출력 h_t를 생성할 수 있다. 이 때, 시점 t의 리커런트 레이어(111)의 출력 x_t에는 가중치 벡터 w_0가 적용되고, 시점 t+1의 리커런트 레이어(111)의 출력 x_{t+1}에는 가중치 벡터 w_1이 적용되며, 시점 t+2의 리커런트 레이어(111)의 출력 x_{t+2}에는 가중치 벡터 w_2가 적용될 수 있다.

[0044] 룩어헤드 컨볼루션 레이어(113)에서 고려되는 미래 시점에서의 리커런트 레이어(111)의 출력을 미래 컨텍스트라고 하며, 룩어헤드 컨볼루션 레이어(113)에서 커버하는 미래 시점의 범위를 미래 컨텍스트 길이라고 지칭할 수 있다.

[0045] 정리하면, 시점 t의 룩어헤드 컨볼루션 레이어(113)의 출력 h_t을 아래와 같이 나타낼 수 있다.

수학식 1

$$h_t = \sum_{j=0}^{\tau} w_j \odot x_{t+j}$$

[0046]

[0047] 위의 수학식 1에서, h_t는 시점 t에서의 룩어헤드 컨볼루션 레이어(113)의 출력을 나타내고, x_{t+j}은 시점 t+j에서의 리커런트 레이어(111)의 출력을 나타내고, w_j은 시점 t+j에서의 리커런트 레이어(111)의 출력에 적용되는 가중치 벡터를 나타낸다.

[0048] 도 3은 일실시예에 따라 스케일링 팩터가 메인 뉴럴 네트워크에 적용되는 과정을 나타낸 도면이다.

[0049] 도 3을 참조하면, 일실시예에 따른 메인 뉴럴 네트워크(110)는 리커런트 레이어(111) 및 룩어헤드 컨볼루션 레이어(113)를 포함하며, 리커런트 레이어(111)의 출력에 스케일링 팩터가 더 적용될 수 있다.

[0050] 일실시예에 따른 스케일링 팩터는 도 2에서 설명한 가중치 벡터와 함께 리커런트 레이어(111)의 출력에 적용될 수 있다. 예를 들어, 시점 t의 리커런트 레이어(111)의 출력 x_t에는 가중치 벡터 w_0뿐만 아니라 스케일링 팩터 내 1차원 성분 α_0도 적용되어, 룩어헤드 컨볼루션 레이어(113)로 전달될 수 있다. 마찬가지로, 시점 t+1의 리커런트 레이어(111)의 출력 x_{t+1}에는 가중치 벡터 w_1과 스케일링 팩터 내 2차원 성분 α_1가 모두 적용되고, 시점 t+2의 리커런트 레이어(111)의 출력 x_{t+2}에는 가중치 벡터 w_2과 스케일링 팩터 내 3차원 성분 α_2가 모두 적용되어 룩어헤드 컨볼루션 레이어(113)로 전달될 수 있다.

[0051] 정리하면, 스케일링 팩터가 적용된 시점 t의 룩어헤드 컨볼루션 레이어(113)의 출력 h_t을 아래와 같이 나타낼 수 있다.

수학식 2

$$h_t = \sum_{j=0}^{\tau} (\alpha_j w_j) \odot x_{t+j}$$

[0052]

[0053] 위의 수학식 2에서, α_j은 시점 t+j에서의 리커런트 레이어(111)의 출력에 적용되는 스케일링 팩터 내 j차원 성분을 나타낸다.

[0054] 스케일링 팩터를 이용하여 룩어헤드 컨볼루션 레이어(113)에서 고려할 미래 컨텍스트 길이를 조절하는 과정에 대해서는 도 5를 참조하여 설명한다.

[0055] 도 4는 일실시예에 따라 서브 뉴럴 네트워크를 나타낸 도면이다.

[0056] 도 4를 참조하면, 일실시예에 따라 서브 뉴럴 네트워크(120)에서 스케일링 팩터(410)가 결정되는 예시가 도시된다. 도 4에서는 설명의 편의를 위해 서브 뉴럴 네트워크(120) 내 하나의 히든 레이어가 도시되어 있으나, 서브 뉴럴 네트워크(120)는 하나 이상의 히든 레이어를 제한 없이 포함할 수 있다. 또한, 히든 레이어는 하나 이상의 히든 노드를 제한 없이 포함할 수 있다.

[0057] 서브 뉴럴 네트워크(120)에서 인접한 레이어에 속한 노드들은 시냅스를 통해 서로 연결될 수 있고, 시냅스에는 미리 트레이닝된 가중치가 부여될 수 있다.

[0058] 일실시예에 따른 서브 뉴럴 네트워크(120)는 도 3에서 설명한 메인 뉴럴 네트워크(110)와 구별되는 뉴럴 네트워크로서, 예를 들어, 리커런트 뉴럴 네트워크, 컨볼루셔널 뉴럴 네트워크, 또는 일반적인 딥 뉴럴 네트워크일 수 있다. 예를 들어, 서브 뉴럴 네트워크(120)는 음성 신호에 대한 컨텍스트 모델(context model)일 수 있다.

[0059] 서브 뉴럴 네트워크(120)는 인식의 대상이 되는 입력을 수신하여 해당 입력에 대응하는 스케일링 팩터(410)를 출력할 수 있다. 예를 들어, 서브 뉴럴 네트워크(120)는 인식의 대상이 되는 음성 신호를 수신하고, 해당 음성 신호에 대한 스케일링 팩터(410)를 출력할 수 있다.

[0060] 일실시예에 따른 스케일링 팩터(410)는 도 3의 메인 뉴럴 네트워크(110)의 리커런트 레이어(111)로부터 룩어헤드 컨볼루션 레이어(113)로 연결되는 가중치 벡터를 스케일링하는 요소로서, n차원 벡터를 포함할 수 있다. 도 4에서는 설명의 편의를 위해 스케일링 팩터(410)가 3차원 벡터에 해당하는 것으로 도시되어 있으나, 스케일링 팩터(410)는 제한 없이 하나 이상의 차원 벡터를 가질 수 있다.

[0061] 일실시예에 따른 스케일링 팩터(410) 내 각 성분들은 해당 성분에 대응하는 컨텍스트의 중요도를 나타낼 수 있다. 예를 들어, 스케일링 팩터(410) 내 1차원 성분 α_0은 시점 t의 리커런트 레이어(111)의 출력에 적용됨으로써, 시점 t의 컨텍스트에 대한 중요도를 나타낼 수 있다. 마찬가지로, 2차원 성분 α_1은 시점 t의 리커런트 레이어(111)의 출력에 적용되어 시점 t+1의 컨텍스트에 대한 중요도를 나타내고, 3차원 성분 α_2은 시점 t+2의 리커런트 레이어(111)의 출력에 적용되어 시점 t+2의 컨텍스트에 대한 중요도를 나타낼 수 있다.

[0062] 서브 뉴럴 네트워크(120)에서 출력되는 스케일링 팩터(410)는 음성 신호가 인식될 때 메인 뉴럴 네트워크에서 고려될 미래 컨텍스트의 시점별 중요도를 나타낼 수 있다. 스케일링 팩터(410)는 인식의 대상이 되는 음성 신호의 명확성에 기초하여 결정될 수 있다. 예를 들어, 스케일링 팩터(410)는 음성 신호에 포함된 노이즈 정도 및 음성 신호에 포함된 사용자 발음의 정확도 중 적어도 하나에 기초하여 결정될 수 있다.

[0063] 만약 음성 신호에 포함된 노이즈 정도가 크거나, 음성 신호에 포함된 사용자 발음의 정확도가 낮은 경우(예컨대, 사용자가 명확히 발음하지 않고, 발음을 흘리는 경우), 스케일링 팩터(410)는 미리 결정된 임계치보다 큰 값을 가지는 성분이 많아지도록 결정될 수 있다. 예를 들어, 음성 신호의 명확성이 낮을수록 스케일링 팩터(410)에 포함된 저차원 성분부터 순차적으로 미리 결정된 임계치보다 큰 값을 가지도록 결정될 수 있다. 따라서, 음성 신호의 명확성이 상당히 낮은 경우, 스케일링 팩터(410)에 포함된 저차원 성분뿐만 아니라 고차원 성분도 미리 결정된 임계치보다 큰 값을 가지도록 결정될 수 있다.

[0064] 반대로, 음성 신호에 포함된 노이즈 정도가 작거나, 음성 신호에 포함된 사용자 발음의 정확도가 높은 경우, 스케일링 팩터(410)는 미리 결정된 임계치보다 큰 값을 가지는 성분이 적어지도록 결정될 수 있다. 예를 들어, 음성 신호의 명확성이 높을수록 스케일링 팩터(410)에 포함된 고차원 성분부터 순차적으로 미리 결정된 임계치보다 작은 값을 가지도록 결정될 수 있다. 따라서, 음성 신호의 명확성이 상당히 높은 경우, 스케일링 팩터(410)에 포함된 고차원 성분뿐만 아니라 저차원 성분도 미리 결정된 임계치보다 작은 값을 가지도록 결정될 수 있다. 다만, 이러한 경우에도 스케일링 팩터(410)의 1차원 성분 α_0은 미리 결정된 임계치보다 큰 값을 가지도록 결정되어, 동일 시점의 컨텍스트가 고려되도록 할 필요가 있다.

[0065] 도 5는 일실시예에 따라 스케일링 팩터에 기초하여 미래 컨텍스트 길이가 결정되는 과정을 나타낸 도면이다.

[0066] 도 5를 참조하면, 일실시예에 따른 메인 뉴럴 네트워크(110)에 포함된 룩어헤드 컨볼루션 레이어(113)의 미래 컨텍스트 길이가 스케일링 팩터에 기초하여 조절되는 예시가 도시된다.

[0067] 도 4의 서브 뉴럴 네트워크(120)에서 결정된 스케일링 팩터의 성분들을 미리 결정된 임계치 ϵ와 비교함으로써, 메인 뉴럴 네트워크(110)에 적용될 미래 컨텍스트 길이를 결정할 수 있다.

[0068] 예를 들어, 스케일링 팩터에 포함된 성분들 중 3차원 성분 α_2이 미리 결정된 임계치 ϵ보다 낮은 값을 가지고, 나머지 성분들 α_0, α_1이 모두 임계치 ϵ보다 큰 값을 가지는 경우, 미래 컨텍스트 길이가 1로 결정되어 시점 t의 룩어헤드 컨볼루션 레이어(113)의 출력을 결정할 때 시점 t+2의 리커런트 레이어(111)의 출력이

배제될 수 있다.

[0069] 이와 같이, 미리 결정된 임계치 ϵ 보다 큰 값을 가지는 스케일링 팩터의 성분들 중 가장 높은 차원의 성분에 기초하여 미래 컨텍스트 길이가 결정될 수 있다. 예를 들어, 인식 장치는 스케일링 팩터에 포함된 복수의 성분들을 고차원 성분부터 미리 결정된 임계치 ϵ 와 비교하고, 미리 결정된 임계치 ϵ 보다 큰 값을 가지는 것으로 처음 확인된 성분의 차원에 기초하여 미래 컨텍스트 길이를 결정할 수 있다.

[0070] 일실시예에 따른 미리 결정된 임계치 ϵ 는 미래 컨텍스트 길이를 결정하는 데 기준이 되는 값으로서, 실험적으로 미리 결정될 수 있다.

[0071] 인식 장치는 스케일링 팩터에 기초하여 메인 뉴럴 네트워크(110)에 포함된 룩어헤드 컨볼루션 레이어(113)의 미래 컨텍스트 길이를 적응적으로 조절함으로써, 최소한의 미래 컨텍스트를 이용하여 높은 정확도와 빠른 반응속도를 기대해 볼 수 있다.

[0072] 도 6은 일실시예에 따라 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크를 트레이닝하는 과정을 나타낸 도면이다.

[0073] 도 6을 참조하면, 일실시예에 따른 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크는 동시에 트레이닝될 수 있다.

[0074] 일실시예에 따른 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크는 트레이닝 데이터에 기초하여 트레이닝될 수 있다. 트레이닝 데이터는 트레이닝 입력 및 트레이닝 출력을 포함할 수 있다. 트레이닝 출력은 트레이닝 입력에 매핑된 출력으로서, 예를 들어, 트레이닝 입력으로부터 출력되어야 하는 레이블(label)일 수 있다. 예를 들어, 음성 인식에 있어서, 트레이닝 입력은 음성 신호이고, 트레이닝 출력은 해당 음성 신호가 나타내는 음소 정보일 수 있다.

[0075] 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크는 역전파 학습(back propagation learning)(610)을 통해, 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크의 레이어 간의 연결 가중치, 노드의 상태 파라미터 등을 트레이닝시킬 수 있다.

[0076] 예를 들어, 메인 뉴럴 네트워크 및 서브 뉴럴 네트워크는 감독 학습(supervised learning)을 통해 트레이닝될 수 있다. 감독 학습이란 트레이닝 입력과 그에 대응하는 트레이닝 출력을 함께 신경 망에 입력하고, 트레이닝 입력에 대응하는 트레이닝 출력이 출력되도록 연결선들의 연결 가중치를 업데이트하는 방법이다.

[0077] 역전파 학습(610)은, 주어진 트레이닝 데이터에 대해 전방 계산(forward computation)으로 손실(loss)을 추정한 후, 출력 레이어에서 시작하여 히든 레이어와 입력 레이어로의 역 방향으로 추정한 손실을 전파하고, 손실을 줄이는 방향으로 연결 가중치를 조절하는 방법이다. 신경 망의 인식을 위한 처리는 입력 레이어, 히든 레이어, 출력 레이어의 순서로 진행되지만, 역전파 학습(610)에서 연결 가중치의 업데이트 방향은 출력 레이어, 히든 레이어, 입력 레이어의 순서로 진행될 수 있다.

[0078] 이 때, 서브 뉴럴 네트워크에서 출력된 스케일링 팩터에 기초하여 룩어헤드 컨볼루션 레이어의 미래 컨텍스트 길이가 조절된 상태로 역전파 학습(610)이 진행될 수 있다. 예를 들어, 스케일링 팩터의 3차원 성분 α_2이 미리 결정된 임계치보다 낮은 값을 가지고, 2차원 성분 α_1이 미리 결정된 임계치보다 큰 값을 가지는 경우, 룩어헤드 컨볼루션 레이어의 미래 컨텍스트 길이는 1로 조절되어 트레이닝이 수행될 수 있다. 이를 통해, 인식을 수행하는 메인 뉴럴 네트워크뿐만 아니라 스케일링 팩터를 출력하는 서브 뉴럴 네트워크도 함께 학습될 수 있다.

[0079] 일실시예에 따른 서브 뉴럴 네트워크는 출력되는 스케일링 팩터의 각 성분들이 0에 가까운 값을 가질 수 있도록 regularizer를 통해 트레이닝될 수 있다. 이를 통해, 스케일링 팩터에 포함된 복수의 성분들 중 고차원 성분일수록 0에 가까운 값을 가지도록 서브 뉴럴 네트워크가 트레이닝될 수 있다.

[0080] 도 7은 일실시예에 따라 스케일링 팩터를 설명하기 위한 도면이다.

[0081] 도 7을 참조하면, 일실시예에 따른 제1 스케일링 팩터(710) 및 제2 스케일링 팩터(720)가 도시된다.

[0082] 제1 스케일링 팩터(710) 및 제2 스케일링 팩터(720)는 서로 다른 음성 신호에 대응하여 서브 뉴럴 네트워크에서 출력된 n차원 벡터를 가질 수 있다. 예를 들어, 제1 스케일링 팩터(710)는 제2 스케일링 팩터(720)보다 높은 명확성을 가지는 음성 신호에 대응될 수 있다. 음성 신호의 명확성이 높다는 것은 해당 음성 신호에 포함된 노이즈 정도가 작거나, 해당 음성 신호에 포함된 사용자 발음의 정확도가 높다는 것을 의미하고, 이 경우 미래 컨

텍스트를 적게 고려하더라도 높은 정확도로 음성 인식이 수행될 수 있다.

[0083] 이와 같이 음성 신호의 명확성에 기초하여 결정된 제1 스케일링 팩터(710)는 제2 스케일링 팩터(720)보다 대체로 작은 값을 가질 수 있다. 제1 스케일링 팩터(710)와 제2 스케일링 팩터(720)는 미리 결정된 임계치 ϵ와의 비교를 통해, 메인 뉴럴 네트워크 내 룩어헤드 컨볼루션 레이어에서 고려할 미래 컨텍스트 길이가 결정될 수 있다.

[0084] 예시적으로 도시된 도 7의 제1 스케일링 팩터(710)의 경우, a+1차원 성분은 미리 결정된 임계치 ϵ보다 낮은 값을 가지며, 1차원 성분부터 a차원 성분은 미리 결정된 임계치 ϵ보다 큰 값을 가진다. 따라서, 제1 스케일링 팩터(710)의 경우, 미래 컨텍스트 길이는 a-1로 결정될 수 있다. 또한, 제2 스케일링 팩터(720)의 경우, b+1차원 성분은 미리 결정된 임계치 ϵ보다 낮은 값을 가지며, 1차원 성분부터 b차원 성분은 미리 결정된 임계치 ϵ보다 큰 값을 가진다. 따라서, 제2 스케일링 팩터(720)의 경우, 미래 컨텍스트 길이는 b-1로 결정될 수 있다.

[0085] 도 8은 일실시예에 따라 스케일링 팩터를 결정하는 과정을 나타낸 도면이다.

[0086] 도 8을 참조하면, 일실시예에 따른 음성 신호의 윈도우를 기준으로 스케일링 팩터(820)가 결정되는 예시가 도시된다.

[0087] 일실시예에 따라 음성 신호는 연속된 일련의 시퀀스 데이터로서, 미리 결정된 길이의 윈도우로 구분되어 인식될 수 있다. 윈도우 길이는 설계에 따라 달리 설정될 수 있다. 예를 들어, 음성 신호가 입력되면 이를 200 msec의 윈도우로 분할하여 음성 인식이 수행될 수 있다.

[0088] 이러한 윈도우를 기준으로 서브 뉴럴 네트워크를 통해 스케일링 팩터(820)도 결정되는 데, 이러한 스케일링 팩터(820)는 윈도우마다 결정될 수 있다. 해당 윈도우에서 결정된 스케일링 팩터(820)에 따라 룩어헤드 컨볼루션 레이어의 미래 컨텍스트 길이를 조절하여 해당 윈도우에 대한 인식 또는 트레이닝이 수행될 수 있다.

[0089] 다른 일실시예에 따르면, 스케일링 팩터(820)는 인식의 대상이 되는 음성 신호의 첫 번째 윈도우(810)에 대해서 결정되고, 첫 번째 윈도우(810)에 대해 결정된 스케일링 팩터(820)에 기초하여 조절된 미래 컨텍스트 길이를 가지는 룩어헤드 컨볼루션 레이어를 이용하여 전체 윈도우에 대한 인식이 수행될 수도 있다. 이는 동일한 음성 신호 내에서는 노이즈 정도나 사용자 발음의 정확도가 크게 달라지지 않는다는 것에 착안할 수 있다.

[0090] 예를 들어, 반응속도에 상대적으로 덜 민감한 트레이닝의 경우, 윈도우마다 스케일링 팩터(820)를 결정할 수 있으나, 반응속도에 민감한 인식의 경우, 음성 신호의 첫 번째 윈도우(810)에 대해 결정한 스케일링 팩터(820)를 이용하여 전체 윈도우에 대해 인식을 수행할 수 있다.

[0091] 도 9는 일실시예에 따른 인식 방법을 나타낸 도면이다.

[0092] 도 9를 참조하면, 일실시예에 따른 인식 장치의 프로세서에서 수행되는 인식 방법이 도시된다.

[0093] 단계(910)에서, 인식 장치는 서브 뉴럴 네트워크를 이용하여 음성 신호로부터 스케일링 팩터를 획득한다. 스케일링 팩터는 음성 신호가 인식될 때 메인 뉴럴 네트워크에서 고려될 미래 컨텍스트의 시점별 중요도를 나타낼 수 있다. 예를 들어, 스케일링 팩터는 음성 신호에 포함된 노이즈 정도 및 음성 신호에 포함된 사용자 발음의 정확도 중 적어도 하나에 기초하여 결정될 수 있다. 스케일링 팩터는 τ 차원 벡터를 포함할 수 있다.

[0094] 단계(920)에서, 인식 장치는 스케일링 팩터 내 가장 높은 차원 성분부터 먼저 고려할 수 있다. 인식 장치는 고려하고자 하는 j차원을 τ로 설정할 수 있다.

[0095] 단계(930)에서, 인식 장치는 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 작은지 여부를 판단할 수 있다.

[0096] 만약 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 작은 경우, 단계(940)에서 인식 장치는 메인 뉴럴 네트워크에서 가중치 벡터 w_j를 제거할 수 있다. 가중치 벡터 w_j는 시점 t의 룩어헤드 컨볼루션 레이어로 전달되는 시점 t+j의 리커런트 레이어의 출력에 적용되는 가중치일 수 있다. 다시 말해, 메인 뉴럴 네트워크에서 가중치 벡터 w_j를 제거함으로써, 시점 t+j의 미래 컨텍스트는 고려에서 제외될 수 있다.

[0097] 단계(950)에서, 인식 장치는 다음에 고려할 차원을 단계(930)에서 판단한 차원보다 한 차원 낮게 설정할 수 있다.

[0098] 그리고, 단계(930)에서, 인식 장치는 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 작은지 여부를 판단할 수 있다. 미리 결정된 임계치 ϵ보다 큰 값을 가지는 j차원 성분 α_j이 확인될 때까지, 단계(930) 내지 단계(950)이 반복해서 수행될 수 있다.

[0099] 만약 단계(930)에서 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 큰 경우, 단계(960)에서 인식 장치는 메인 뉴럴 네트워크에 포함된 룩어헤드 컨볼루션 레이어의 미래 컨텍스트 길이를 j로 조절할 수 있다.

[0100] 단계(970)에서, 인식 장치는 미래 컨텍스트 길이가 j로 조절된 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크를 이용하여 음성 신호를 인식할 수 있다.

[0101] 도 9에 도시된 각 단계들에는 도 1 내지 도 8을 통하여 전술한 사항들이 마찬가지로 적용되므로, 보다 상세한 설명은 생략한다.

[0102] 도 10은 일실시예에 따른 트레이닝 방법을 나타낸 도면이다.

[0103] 도 10을 참조하면, 일실시예에 따른 트레이닝 장치의 프로세서에서 수행되는 트레이닝 방법이 도시된다.

[0104] 단계(1010)에서, 트레이닝 장치는 서브 뉴럴 네트워크를 이용하여 트레이닝 입력으로부터 스케일링 팩터를 획득한다.

[0105] 단계(1020)에서, 트레이닝 장치는 스케일링 팩터 내 가장 높은 차원 성분부터 먼저 고려할 수 있다. 트레이닝 장치는 고려하고자 하는 j차원을 τ로 설정할 수 있다.

[0106] 단계(1030)에서, 트레이닝 장치는 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 작은지 여부를 판단할 수 있다.

[0107] 만약 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 작은 경우, 단계(1040)에서, 트레이닝 장치는 메인 뉴럴 네트워크에서 가중치 벡터 w_j를 제거할 수 있다. 가중치 벡터 w_j는 시점 t의 룩어헤드 컨볼루션 레이어로 전달되는 시점 t+j의 리커런트 레이어의 출력에 적용되는 가중치일 수 있다. 다시 말해, 메인 뉴럴 네트워크에서 가중치 벡터 w_j를 제거함으로써, 시점 t+j의 미래 컨텍스트는 고려에서 제외될 수 있다.

[0108] 단계(1050)에서, 트레이닝 장치는 다음에 고려할 차원을 단계(1030)에서 판단한 차원보다 한 차원 낮게 설정할 수 있다.

[0109] 그리고, 단계(1030)에서, 트레이닝 장치는 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 작은지 여부를 판단할 수 있다. 미리 결정된 임계치 ϵ보다 큰 값을 가지는 j차원 성분 α_j이 확인될 때까지, 단계(1030) 내지 단계(1050)이 반복해서 수행될 수 있다.

[0110] 만약 단계(1030)에서 스케일링 팩터의 j차원 성분 α_j이 미리 결정된 임계치 ϵ보다 큰 경우, 트레이닝 장치는 메인 뉴럴 네트워크에 포함된 룩어헤드 컨볼루션 레이어의 미래 컨텍스트 길이를 j로 조절할 수 있다.

[0111] 단계(1060)에서, 트레이닝 장치는 미래 컨텍스트 길이가 j로 조절된 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크와 서브 뉴럴 네트워크를 트레이닝시킬 수 있다. 예를 들어, 트레이닝 장치는 미래 컨텍스트 길이가 j로 조절된 컨볼루션 레이어를 포함한 메인 뉴럴 네트워크와 서브 뉴럴 네트워크를 이용하여 트레이닝 입력으로부터 트레이닝 출력이 출력되도록 메인 뉴럴 네트워크와 서브 뉴럴 네트워크를 동시에 트레이닝시킬 수 있다.

[0112] 도 10에 도시된 각 단계들에는 도 1 내지 도 9을 통하여 전술한 사항들이 마찬가지로 적용되므로, 보다 상세한 설명은 생략한다.

[0113] 도 11은 일실시예에 따른 인식 장치를 나타낸 도면이다.

[0114] 도 11을 참조하면, 일실시예에 따른 인식 장치(1100)는 프로세서(1110) 및 메모리(1120)를 포함한다.

[0115] 메모리(1120)는 앞서 설명한 메인 뉴럴 네트워크(1121) 및 서브 뉴럴 네트워크(1123)의 파라미터들을 저장할 수 있다. 또한, 메모리(1120)는 컴퓨터에서 읽을 수 있는 명령어를 포함할 수 있다. 프로세서(1110)는 메모리(1120)에 저장된 명령어가 프로세서(1110)에서 실행됨에 따라 앞서 언급된 동작들을 수행할 수 있다. 메모리(1120)는 휘발성 메모리 또는 비휘발성 메모리일 수 있다.

[0116] 프로세서(1110)는 메모리(1120)에서 메인 뉴럴 네트워크(1121) 및 서브 뉴럴 네트워크(1123)에 관련된 데이터를 획득하고, 메인 뉴럴 네트워크(1121) 및 서브 뉴럴 네트워크(1123)에 관련된 동작을 처리할 수 있다.

[0117] 일실시예에 따른 프로세서(1110)는 음성 신호가 입력된 서브 뉴럴 네트워크(1123)로부터 스케일링 팩터를 획득하고, 스케일링 팩터에 기초하여 메인 뉴럴 네트워크(1121)에서 고려할 미래 컨텍스트 길이를 결정하며, 결정된 미래 컨텍스트 길이가 적용된 메인 뉴럴 네트워크(1121)에 음성 신호가 입력됨에 따라 메인 뉴럴 네트워크(1121)에서 출력되는 음성 신호의 인식 결과를 획득한다.

[0118] 그 밖에, 인식 장치(1100)에는 전술된 사항이 적용될 수 있으며, 보다 상세한 설명은 생략한다.

[0119] 도 12는 일실시예에 따른 트레이닝 장치를 나타낸 도면이다.

[0120] 도 12를 참조하면, 일실시예에 따른 트레이닝 장치(120)는 프로세서(1210) 및 메모리(1220)를 포함한다.

[0121] 메모리(1220)는 앞서 설명한 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)의 파라미터들을 저장할 수 있다. 또한, 메모리(1220)는 컴퓨터에서 읽을 수 있는 명령어를 포함할 수 있다.

[0122] 프로세서(1210)는 메모리(1220)에 저장된 명령어가 프로세서(1210)에서 실행됨에 따라 앞서 언급된 동작들을 수행할 수 있다. 프로세서(1210)는 메모리(1220)에서 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)를 획득할 수 있다. 프로세서(1210)는 트레이닝 데이터(1201)에 기초하여 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)를 트레이닝시킬 수 있다.

[0123] 트레이닝 데이터(1201)는 트레이닝 입력 및 트레이닝 출력을 포함할 수 있다. 트레이닝 입력은 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)에 입력되는 입력 데이터로, 예컨대 음성 데이터를 포함할 수 있다. 트레이닝 출력은 트레이닝 입력에 매핑된 데이터로, 예컨대 트레이닝 입력이 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)에 입력됨에 따라 메인 뉴럴 네트워크(1221)에서 출력되어야 하는 레이블일 수 있다.

[0124] 프로세서(1210)는 트레이닝 입력으로부터 트레이닝 출력이 생성되도록 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)를 트레이닝시킬 수 있다. 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)를 훈련시킨다는 것은 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)의 파라미터를 훈련시키는 것, 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)를 갱신하는 것, 혹은 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)의 파라미터를 갱신하는 것을 포함할 수 있다.

[0125] 프로세서(1210)는 트레이닝 입력이 입력된 서브 뉴럴 네트워크(1223)로부터 스케일링 팩터를 획득하고, 스케일링 팩터에 기초하여 메인 뉴럴 네트워크(1221)에서 고려할 미래 컨텍스트 길이를 결정한다. 그리고, 프로세서(1210)는 결정된 미래 컨텍스트 길이가 적용된 메인 뉴럴 네트워크(1221)에 트레이닝 입력이 입력됨에 따라 메인 뉴럴 네트워크(1221)에서 트레이닝 입력에 매핑된 트레이닝 출력이 출력되도록, 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)를 트레이닝시킨다.

[0126] 메인 뉴럴 네트워크(1221)는 서브 뉴럴 네트워크(1223)의 출력(예컨대, 스케일링 팩터)에 기초하여 동작하므로, 메인 뉴럴 네트워크(1221) 및 서브 뉴럴 네트워크(1223)는 동시에 트레이닝될 수 있다.

[0127] 그 밖에, 트레이닝 장치(1200)에는 전술된 사항이 적용될 수 있으며, 보다 상세한 설명은 생략한다.

[0128] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나

이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

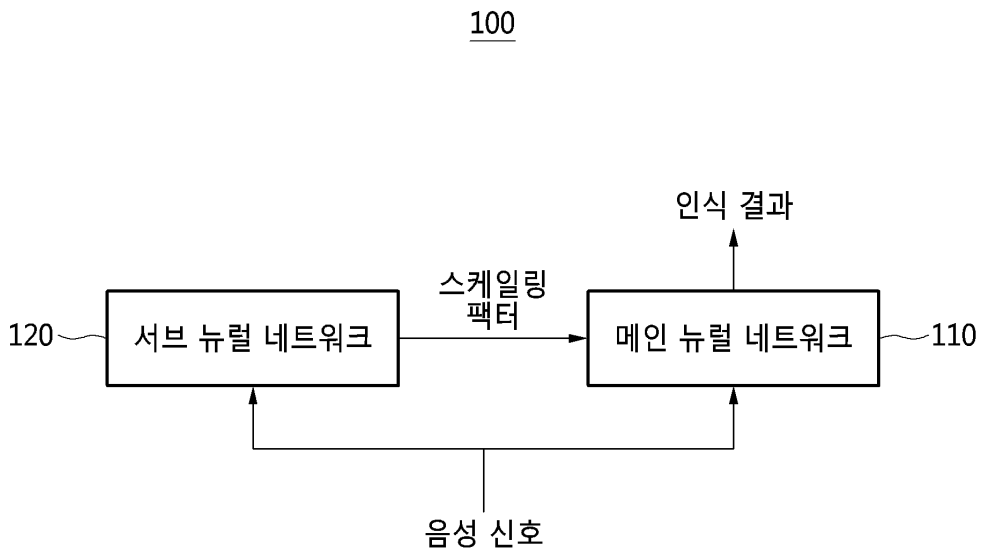
[0129] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 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.

[0130] 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.

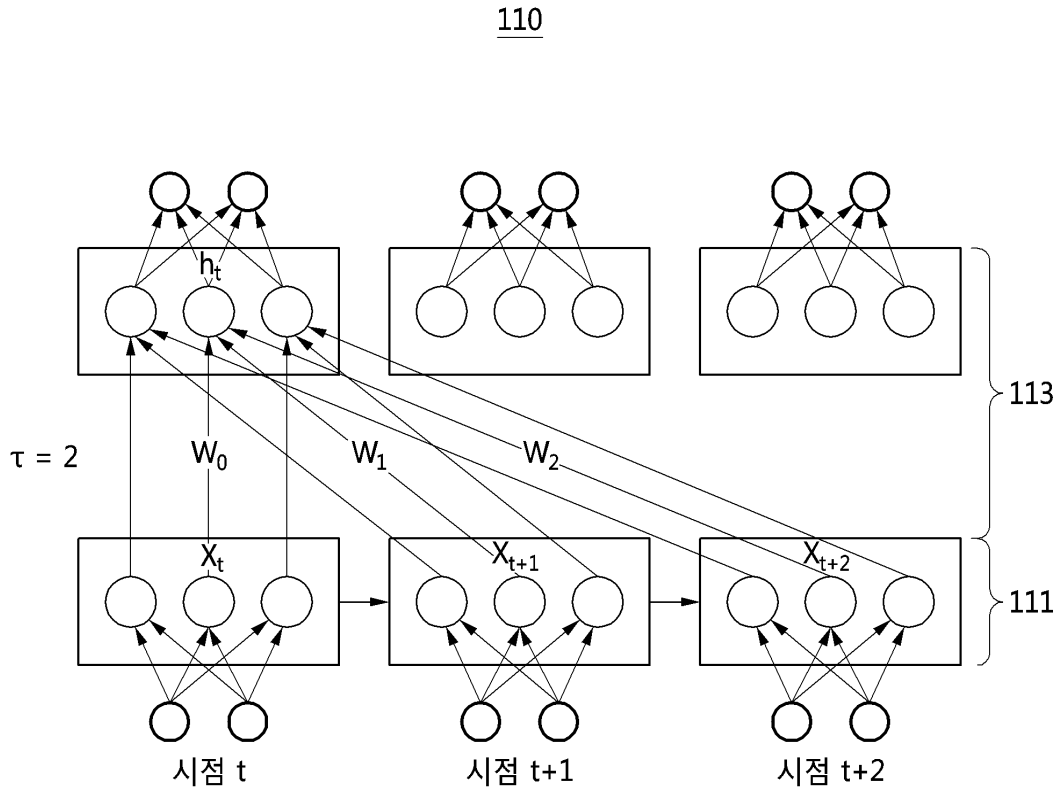
[0131] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

도면

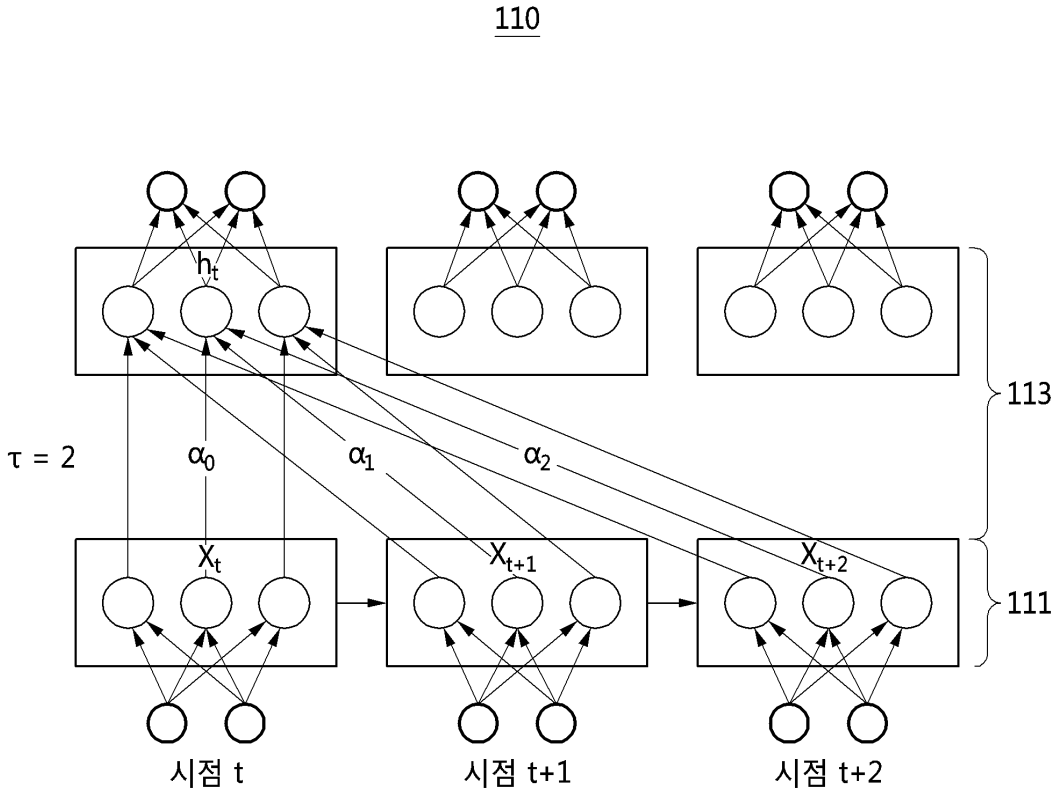
도면1



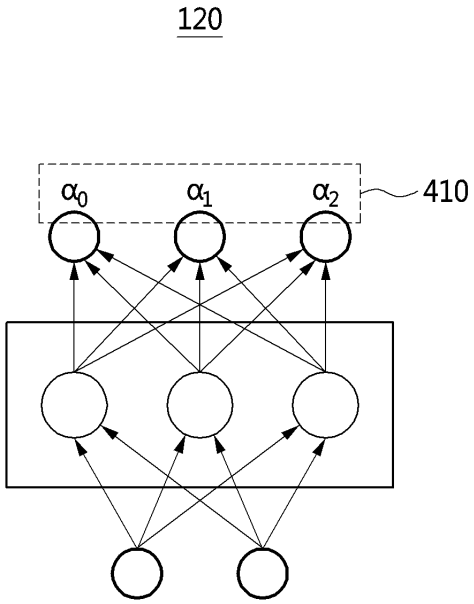
도면2


110
h_t
τ = 2
W_0
W_1
W_2
113
X_t
X_{t+1}
X_{t+2}
111
시점 t
시점 t+1
시점 t+2

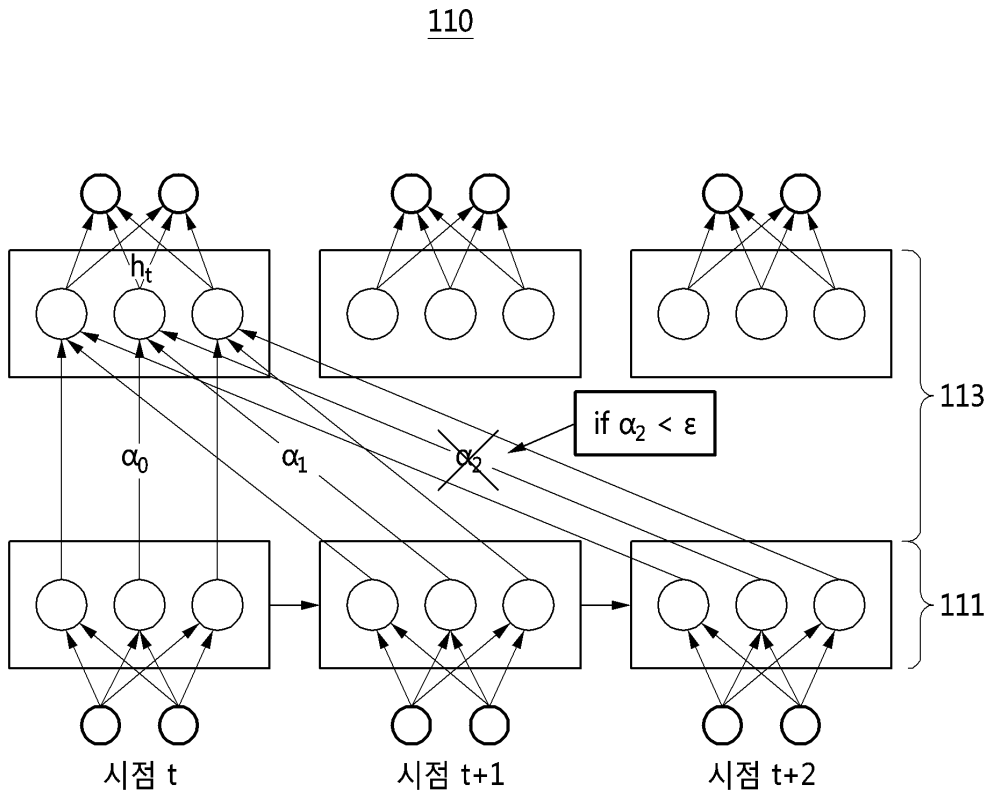
도면3


110
h_t
τ = 2
α_0
α_1
α_2
113
X_t
X_{t+1}
X_{t+2}
111
시점 t
시점 t+1
시점 t+2

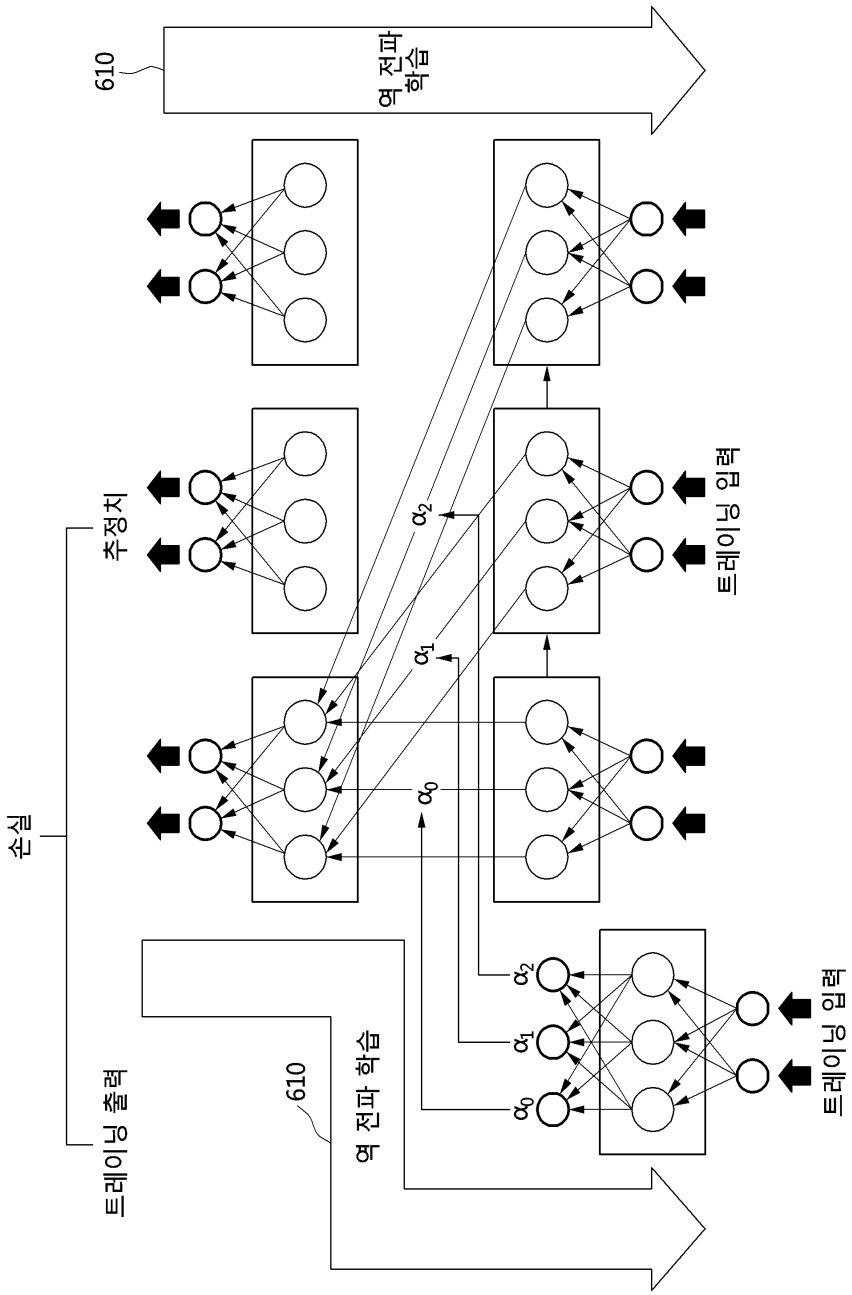
도면4


120
α_0
α_1
α_2
410

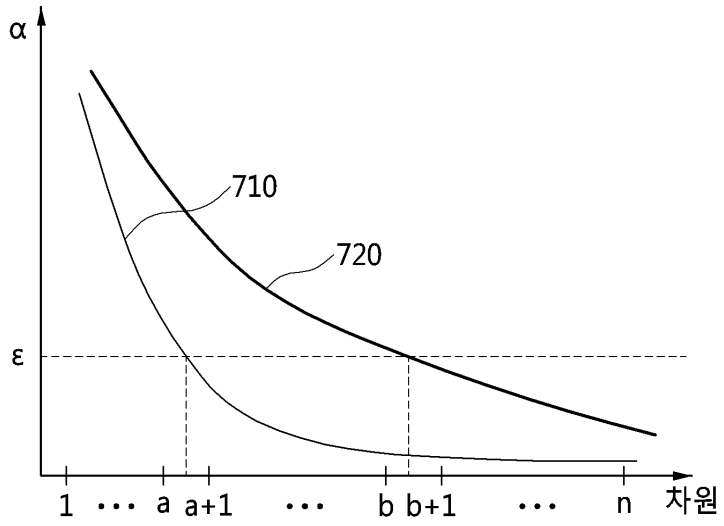
도면5


110
h_t
if $\alpha_2 < \varepsilon$
α_0
α_1
α_2
113
111
시점 t
시점 t+1
시점 t+2

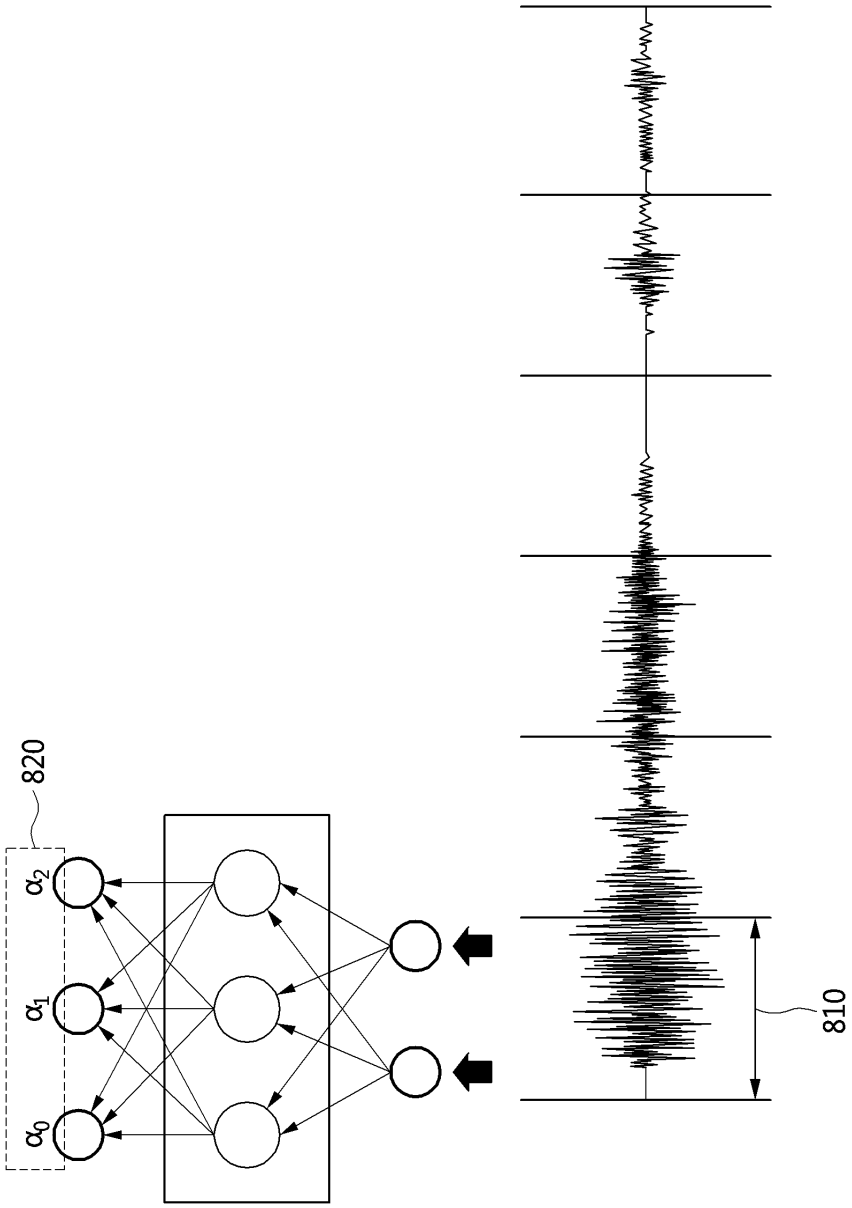
도면6


610
역 전파
학습
추정치
손실
트레이닝 출력
트레이닝 입력
트레이닝 입력
610
역 전파 학습
α_2
α_1
α_0
α_2
α_1
α_0

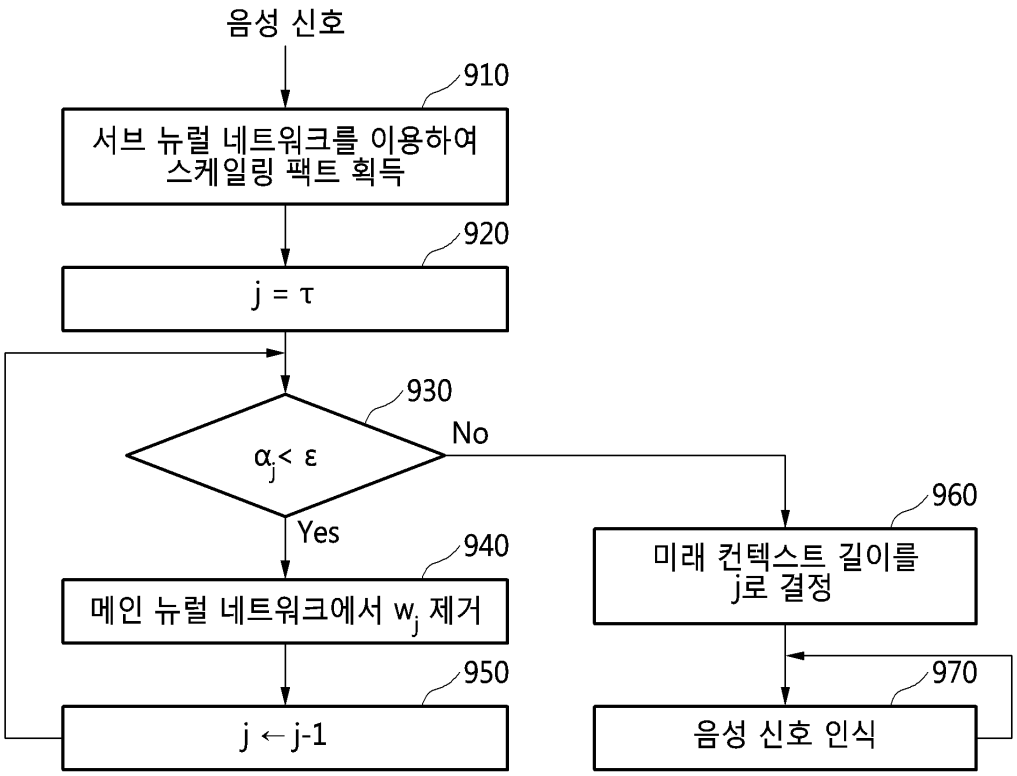
도면7



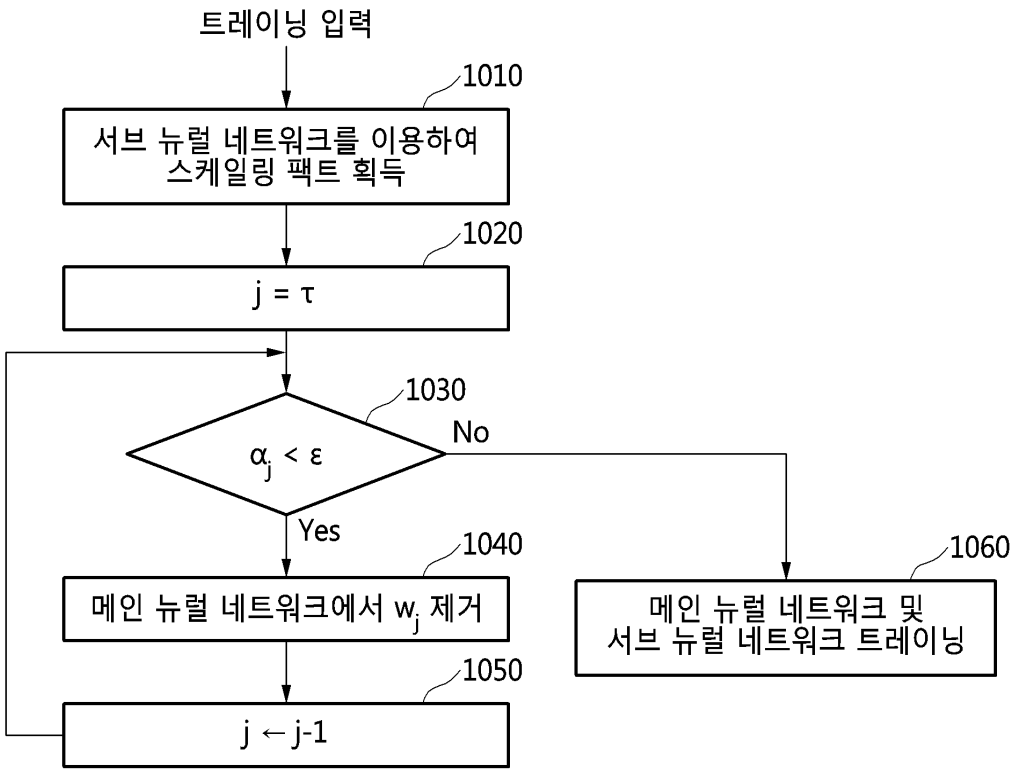
도면8



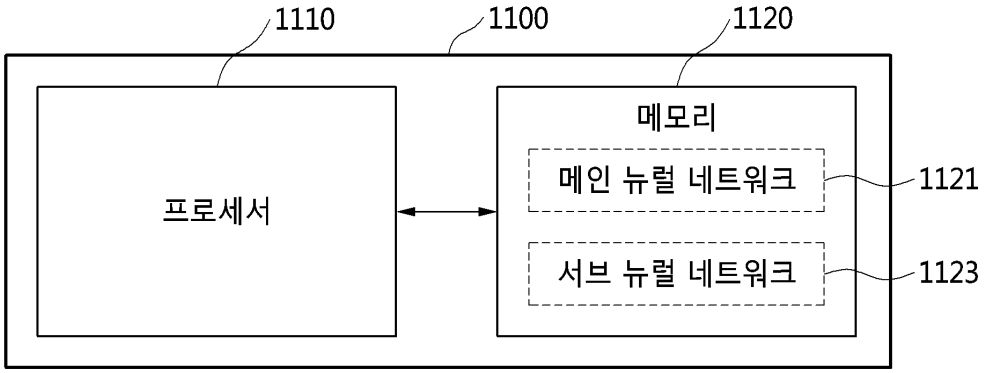
도면9


음성 신호
910
서브 뉴럴 네트워크를 이용하여
스케일링 팩트 획득
920
j = τ
930
$\alpha_j < \varepsilon$
No
Yes
940
메인 뉴럴 네트워크에서 w_j 제거
950
j ← j-1
960
미래 컨텍스트 길이를
j로 결정
970
음성 신호 인식

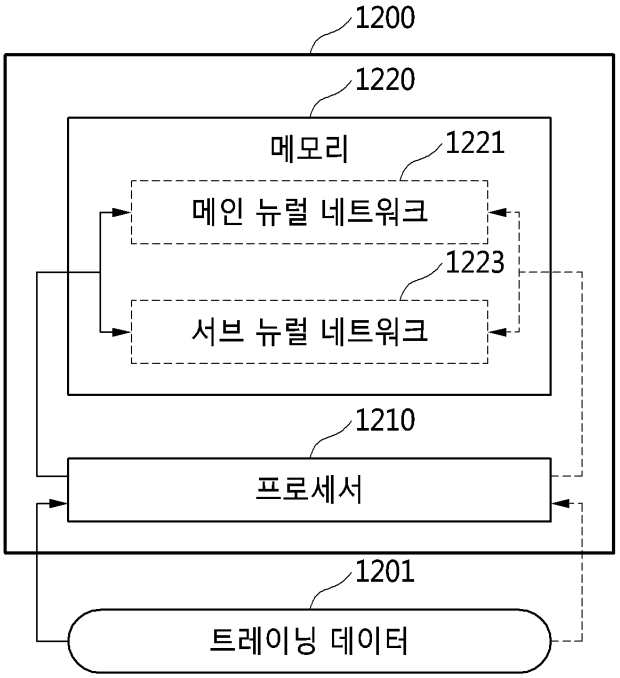
도면10


트레이닝 입력
1010
서브 뉴럴 네트워크를 이용하여
스케일링 팩트 획득
1020
j = τ
1030
$\alpha_j < \varepsilon$
No
Yes
1040
메인 뉴럴 네트워크에서 w_j 제거
1050
j ← j-1
1060
메인 뉴럴 네트워크 및
서브 뉴럴 네트워크 트레이닝

도면11


1110
1100
1120
프로세서
메모리
메인 뉴럴 네트워크
1121
서브 뉴럴 네트워크
1123

도면12


1200
1220
메모리
1221
메인 뉴럴 네트워크
1223
서브 뉴럴 네트워크
1210
프로세서
1201
트레이닝 데이터