(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0086214
(43) 공개일자 2020년07월16일

(51) 국제특허분류(Int. Cl.)
G10L 15/04 (2006.01) *G10L 15/14* (2006.01)
G10L 15/18 (2006.01) *G10L 15/22* (2006.01)
(52) CPC특허분류
G10L 15/04 (2013.01)
G10L 15/14 (2013.01)
(21) 출원번호 10-2019-0151166
(22) 출원일자 2019년11월22일
심사청구일자 2019년11월22일
(30) 우선권주장
201910016861.1 2019년01월08일 중국(CN)

(71) 출원인
바이두 온라인 네트웍 테크놀러지 (베이징) 캄파니 리미티드
중국 베이징 하이뎬 디스트릭트 샹디 10번가 10 바이두 캠퍼스 3층
(72) 발명자
후, 시아오인
중국 베이징 100085 하이뎬 디스트릭트 샹디 10번가 바이두 캠퍼스 3층
바이, 진펑
중국 베이징 100085 하이뎬 디스트릭트 샹디 10번가 바이두 캠퍼스 3층
(뒷면에 계속)
(74) 대리인
이대호, 박건홍

전체 청구항 수 : 총 18 항

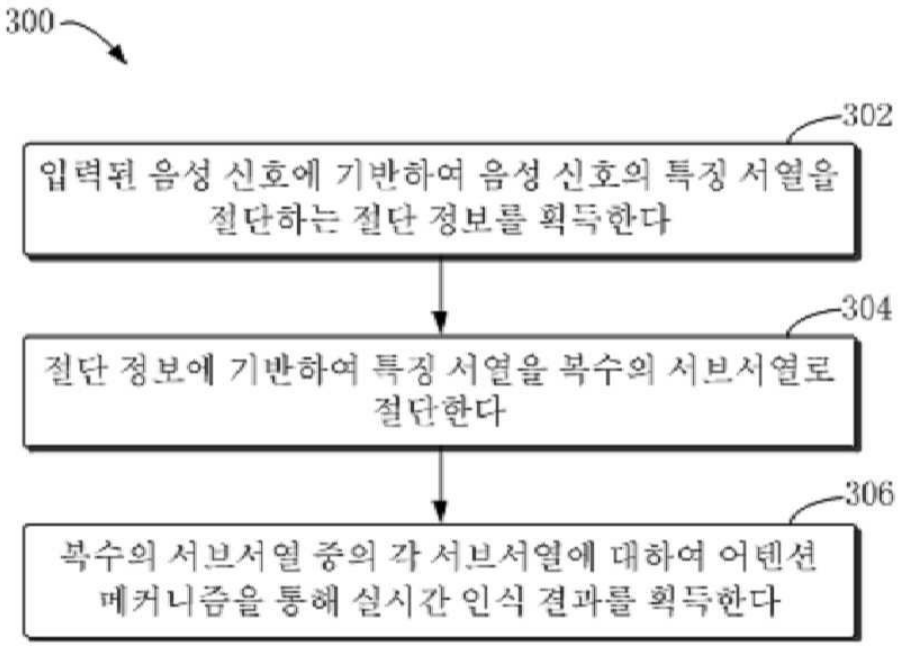
(54) 발명의 명칭 **절단 어텐션에 기반하는 실시간 음성 인식 방법, 장치, 기기 및 컴퓨터 판독 가능 저장 매체**

(57) 요 약

본 개시는 절단 어텐션에 기반하는 실시간 음성 인식 방법, 장치, 기기 및 컴퓨터 판독 가능 저장 매체를 제공한다. 방법은, 입력된 음성 신호의 특징 서열을 절단하는 절단 정보를 획득하고, 절단 정보에 기반하여 특징 서열을 복수의 서브서열로 절단하는 단계를 포함한다. 방법은, 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간

(뒷면에 계속)

대 표 도 - 도3



인식 결과를 획득하는 단계를 더 포함한다. 본 개시의 실시예는 전통적인 어텐션 모델에 음성 신호에 따라 결정되는 절단 정보를 도입함으로써, 각 절단에 대하여 어텐션 모델링을 각각 수행하도록 어텐션 모델을 지도할 수 있는바, 실시간 음성 인식을 구현할 수 있을 뿐만 아니라 매우 높은 인식 정확률을 보장할 수 있다. 또한, 선택적으로, 본 개시의 일부 실시예에 의하여 제공되는 듀얼 헤드 구조는 실시간 음성 인식 과정에서의 계산량이 보다 적도록 보장할 수 있고, 본 개시의 일부 실시예에 의하여 제공되는 복수 등급의 어텐션 구조는 음성 인식의 정확률을 더 향상시킬 수 있다.

(52) CPC특허분류
G10L 15/18 (2013.01)
G10L 15/22 (2013.01)
G10L 2015/221 (2013.01)

(72) 발명자
첸, 쯔지에
중국 베이징 100085 하이뎬 디스트릭트 샹디 10번가 바이두 캠퍼스 3층
량, 밍신
중국 베이징 100085 하이뎬 디스트릭트 샹디 10번가 바이두 캠퍼스 3층
첸, 쑤
중국 베이징 100085 하이뎬 디스트릭트 샹디 10번가 바이두 캠퍼스 3층
지아, 레이
중국 베이징 100085 하이뎬 디스트릭트 샹디 10번가 바이두 캠퍼스 3층

명 세 서

청구범위

청구항 1

절단 어텐션에 기반하는 실시간 음성 인식 방법에 있어서,

입력된 음성 신호에 기반하여, 상기 음성 신호의 특징 서열을 절단하는 절단 정보를 획득하는 단계;

상기 절단 정보에 기반하여, 상기 특징 서열을 복수의 서브서열로 절단하는 단계; 및

상기 복수의 서브서열 중의 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간 인식 결과를 획득하는 단계; 를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 2

제1항에 있어서,

음성 신호의 특징 서열을 절단하는 절단 정보를 획득하는 단계는,

상기 특징 서열에 대하여 연결성 시계열 분류(CTC) 처리를 수행하여, 상기 음성 신호에 관련되는 스파이크 정보를 획득하는 단계; 및

획득된 상기 스파이크 정보에 기반하여 상기 절단 정보를 결정하는 단계; 를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 3

제1항에 있어서,

어텐션 메커니즘을 통해 실시간 인식 결과를 획득하는 단계는,

상기 복수의 서브서열 중의 제1 서브서열에 대하여 어텐션 모델의 제1 어텐션 모델링을 수행하는 단계 - 상기 어텐션 모델은 가중 특징 선택을 구현 가능함 - ; 및

상기 복수의 서브서열 중의 제2 서브서열에 대하여 상기 어텐션 모델의 제2 어텐션 모델링을 수행하는 단계 - 상기 제1 어텐션 모델링과 상기 제2 어텐션 모델링은 상이함 - ; 를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 4

제2항에 있어서,

공유 인코더가 상기 음성 신호를 상기 특징 서열로 인코딩하고;

연결성 시계열 분류(CTC) 모듈이 상기 특징 서열에 기반하여 상기 스파이크 정보를 획득하고;

어텐션 디코더가 상기 특징 서열 및 상기 스파이크 정보에 기반하여, 상기 실시간 인식 결과를 획득하고, 상기 공유 인코더는 상기 연결성 시계열 분류(CTC) 모듈 및 상기 어텐션 디코더에 공유되는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 5

제4항에 있어서,

상기 연결성 시계열 분류(CTC) 모듈을 트레이닝하는 제1 손실 함수를 결정하는 단계;

상기 어텐션 디코더를 트레이닝하는 제2 손실 함수를 결정하는 단계; 및

상기 제1 손실 함수와 상기 제2 손실 함수를 사용하여 상기 연결성 시계열 분류(CTC) 모듈과 상기 어텐션 디코더를 함께 트레이닝하는 단계; 를 더 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 6

제2항에 있어서,

상기 특징 서열을 복수의 서브서열로 절단하는 단계는,

상기 스파이크 정보 중의 각 스파이크에 대하여 상기 특징 서열에서 상기 각 스파이크와 인접하는 소정 개수의 스파이크에 대응되는 서브서열을 선택하는 단계를 포함하고, 상기 소정 개수의 스파이크는 상기 각 스파이크 이전의 제1 개수의 스파이크와 상기 각 스파이크 이후의 제2 개수의 스파이크를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 7

제6항에 있어서,

어텐션 메커니즘을 통해 실시간 인식 결과를 획득하는 단계는,

상기 특징 서열 중의 복수의 중첩되는 서브서열에 기반하여 제1 등급 어텐션 메커니즘을 통해 복수의 특징 벡터를 각각 결정하는 단계;

결정된 상기 복수의 특징 벡터에 기반하여 제2 등급 어텐션 메커니즘을 통해 문맥 특징 벡터를 결정하는 단계; 및

결정된 상기 문맥 특징 벡터에 기반하여 상기 실시간 인식 결과를 결정하는 단계; 를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 8

제6항에 있어서,

상기 특징 서열은 잠재적 특징 서열이고, 상기 각 스파이크는 하나의 음절 또는 한 그룹의 음소를 나타내는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 방법.

청구항 9

절단 어텐션에 기반하는 실시간 음성 인식 장치에 있어서,

입력된 음성 신호에 기반하여, 상기 음성 신호의 특징 서열을 절단하는 절단 정보를 획득하도록 구성되는 절단 정보 획득 모듈;

상기 절단 정보에 기반하여, 상기 특징 서열을 복수의 서브서열로 절단하도록 구성되는 특징 서열 절단 모듈; 및

상기 복수의 서브서열 중의 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간 인식 결과를 획득하도록 구성되는 인식 결과 획득 모듈; 을 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 10

제9항에 있어서,

상기 절단 정보 획득 모듈은,

상기 특징 서열에 대하여 연결성 시계열 분류(CTC) 처리를 수행하여, 상기 음성 신호에 관련되는 스파이크 정보를 획득하도록 구성되는 스파이크 정보 획득 모듈; 및

획득된 상기 스파이크 정보에 기반하여 상기 절단 정보를 결정하도록 구성되는 절단 정보 결정 모듈; 을 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 11

제9항에 있어서,

상기 인식 결과 획득 모듈은,

상기 복수의 서브서열 중의 제1 서브서열에 대하여 어텐션 모델의 제1 어텐션 모델링을 수행하도록 구성되는 제1 어텐션 모델링 장치 - 상기 어텐션 모델은 가중 특징 선택을 구현 가능함 - ; 및

상기 복수의 서브서열 중의 제2 서브서열에 대하여 상기 어텐션 모델의 제2 어텐션 모델링을 수행하도록 구성되는 제2 어텐션 모델링 장치 - 상기 제1 어텐션 모델링과 상기 제2 어텐션 모델링은 상이함 - ; 를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 12

제10항에 있어서,

상기 실시간 음성 인식 장치는,

상기 음성 신호를 상기 특징 서열로 인코딩하도록 구성되는 공유 인코더;

상기 특징 서열에 기반하여 상기 스파이크 정보를 획득하도록 구성되는 연결성 시계열 분류(CTC) 모듈; 및

상기 특징 서열 및 상기 스파이크 정보에 기반하여 상기 실시간 인식 결과를 획득하도록 구성되는 어텐션 디코더 - 상기 공유 인코더는 상기 연결성 시계열 분류(CTC) 모듈 및 상기 어텐션 디코더에 공유됨 - ; 를 포함하고,

상기 절단 정보 획득 모듈은 상기 연결성 시계열 분류(CTC) 모듈에 포함되고, 상기 특징 서열 절단 모듈 및 상기 인식 결과 획득 모듈은 상기 어텐션 디코더에 포함되는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 13

제12항에 있어서,

상기 연결성 시계열 분류(CTC) 모듈을 트레이닝하는 제1 손실 함수를 결정하도록 구성되는 제1 손실 함수 결정 모듈;

상기 어텐션 디코더를 트레이닝하는 제2 손실 함수를 결정하도록 구성되는 제2 손실 함수 결정 모듈; 및

상기 제1 손실 함수와 상기 제2 손실 함수를 사용하여 상기 연결성 시계열 분류(CTC) 모듈과 상기 어텐션 디코더를 함께 트레이닝하도록 구성되는 트레이닝 모듈; 을 더 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 14

제10항에 있어서,

상기 특징 서열 절단 모듈은,

상기 스파이크 정보 중의 각 스파이크에 대하여 상기 특징 서열에서 상기 각 스파이크와 인접하는 소정 개수의 스파이크에 대응되는 서브서열을 선택하도록 구성되는 서브서열 선택 모듈을 포함하고, 상기 소정 개수의 스파이크는 상기 각 스파이크 이전의 제1 개수의 스파이크와 상기 각 스파이크 이후의 제2 개수의 스파이크를 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 15

제14항에 있어서,

상기 인식 결과 획득 모듈은,

상기 특징 서열 중의 복수의 중첩되는 서브서열에 기반하여 제1 등급 어텐션 메커니즘을 통해 복수의 특징 벡터를 각각 결정하도록 구성되는 복수의 특징 벡터 결정 모듈;

결정된 상기 복수의 특징 벡터에 기반하여 제2 등급 어텐션 메커니즘을 통해 문맥 특징 벡터를 결정하도록 구성되는 문맥 특징 벡터 결정 모듈; 및

결정된 상기 문맥 특징 벡터에 기반하여 상기 실시간 인식 결과를 결정하도록 구성되는 인식 결과 결정 모듈;을 포함하는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 16

제14항에 있어서,

상기 특징 서열은 잠재적 특징 서열이고, 상기 각 스파이크는 하나의 음절 또는 한 그룹의 음소를 나타내는,

것을 특징으로 하는 절단 어텐션에 기반하는 실시간 음성 인식 장치.

청구항 17

전자 기기에 있어서,

하나 또는 복수의 프로세서; 및

하나 또는 복수의 프로그램을 저장하는 저장 장치; 를 포함하고,

상기 하나 또는 복수의 프로그램이 상기 하나 또는 복수의 프로세서에 의해 실행되는 경우, 상기 전자 기기가 제1항 내지 제8항 중 어느 한항의 방법을 구현하는,

것을 특징으로 하는 전자 기기.

청구항 18

컴퓨터 프로그램이 저장되어 있는 컴퓨터 판독 가능 저장 매체에 있어서,

상기 프로그램이 프로세서에 의해 실행되는 경우, 제1항 내지 제8항 중 어느 한항의 방법이 실행되는,

것을 특징으로 하는 컴퓨터 판독 가능 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 개시의 실시예는 전체적으로는 음성 인식 기술 분야에 관한 것으로, 더 구체적으로는 절단 어텐션(truncated attention)에 기반하는 실시간 음성 인식 방법, 장치, 기기 및 컴퓨터 판독 가능 저장 매체에 관한 것이다.

배 경 기 술

[0003] 음성 인식은 컴퓨터에 의하여 음성 신호를 대응되는 텍스트로 변환하는 과정을 가리키는 것으로, 인간과 기계의 인터랙션을 구현하는 주요한 경로 중 하나이다. 최근 몇 년 동안, 심층 학습 기술이 음성 인식 분야에서 광범위하게 사용됨에 따라, 음성 인식의 정확률은 매우 크게 향상되었다. 또한, 스마트 기기가 부단히 업그레이드됨으로 인하여 음성을 사용하여 인식되는 장소는 이미 매우 다양화되어 있다. 예를 들면, 음성 인식 기술은 이미 음성 입력, 음성 호출, 차량용 내비게이션 등 다양한 장소에 광범위하게 응용되어 있다. 음성 인식 기술에 자연어

처리 및 음성 합성 등의 기술이 결합되면서 더 많은 복잡한 앱들이 나타날 수 있는바, 예를 들면 스마트 스피커, 회의 동시 통역, 스마트 고객 서비스 보조 등이 있다. 음성 인식의 정확률은 음성 관련 제품 사용자의 사용 경험에 직접 영향 주게 되므로 음성 인식 사용 시나리오가 부단히 다양해짐에 따라, 음성 인식의 정확률에 대해서도 더 높은 요구가 제기되고 있다.

[0004] 실시간 음성 인식은 수신되는 연속 음성에 대하여 음성의 각 세그먼트를 인식하는 것을 가리키는 것으로, 실시간으로 인식 결과를 획득 가능한바, 모든 음성 입력이 끝날 때까지 대기하고 나서 인식을 개시할 필요가 없다. 대규모 단어량에 대한 온라인 연속적인 음성 인식에 있어서 시스템 성능에 영향 주는 핵심 요소는 시스템의 인식 정확도와 응답 속도이다. 예를 들면, 사용자가 말을 하는 동시에 인식 결과가 실시간으로 나타나는 것이 보여지는 것을 희망하는 시나리오에서는 음성 인식 시스템이 높은 인식률을 유지하면서 적시에 빠른 속도로 음성 신호를 디코딩하여 인식 결과를 출력하는 것이 필요하다. 사용자의 음성 입력이 완료된 후에야 인식을 개시하는 것은 사용자의 사용 경험에 크게 영향 주게 되고 음성 인터랙션이 원활하지 않게 된다.

발명의 내용

[0006] 본 개시의 예시적인 실시예에 따르면, 절단 어텐션에 기반하는 실시간 음성 인식 방법, 장치, 기기 및 컴퓨터 판독 가능 저장 매체가 제공된다.

[0007] 본 개시의 제1 측면으로, 절단 어텐션에 기반하는 실시간 음성 인식 방법이 제공된다. 당해 방법은, 입력된 음성 신호에 기반하여 음성 신호의 특징 서열을 절단하는 절단 정보를 획득하는 단계; 절단 정보에 기반하여 특징 서열을 복수의 서브서열로 절단하는 단계; 및 복수의 서브서열 중의 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간 인식 결과를 획득하는 단계; 를 포함한다.

[0008] 본 개시의 제2 측면으로, 절단 어텐션에 기반하는 실시간 음성 인식 장치가 제공된다. 당해 장치는, 입력된 음성 신호에 기반하여 음성 신호의 특징 서열을 절단하는 절단 정보를 획득하도록 구성되는 절단 정보 획득 모듈; 절단 정보에 기반하여 특징 서열을 복수의 서브서열로 절단하도록 구성되는 특징 서열 절단 모듈; 및; 복수의 서브서열 중의 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간 인식 결과를 획득하도록 구성되는 인식 결과 획득 모듈; 을 포함한다.

[0009] 본 개시의 제 3 측면으로, 전자 기기가 제공되는바, 이는 하나 또는 복수의 프로세서 및 하나 또는 복수의 프로그램을 저장하는 저장 장치를 포함한다. 하나 또는 복수의 프로그램이 하나 또는 복수의 프로세서에 의해 실행되는 경우, 전자 기기가 본 개시의 실시예에 따른 방법 또는 과정을 구현한다.

[0010] 본 개시의 제4 측면으로, 컴퓨터 프로그램이 저장되어 있는 컴퓨터 판독 가능 저장 매체가 제공되는바, 당해 프로그램이 프로세서에 의해 구현되는 경우, 본 개시의 실시예에 따른 방법 또는 과정이 구현된다.

[0011] 본 발명 내용 부분에서 설명하는 내용은 본 개시의 실시예의 핵심 특징 또는 중요 특징을 한정하고자 하는 것이 아니며 본 개시의 범위를 제한하기 위한 것도 아님을 이해하여야 한다. 본 개시의 기타 특징은 이하의 설명을 통하여 이해하기 쉽게 될 것이다.

도면의 간단한 설명

[0013] 첨부 도면을 결부하고 이하 상세한 설명을 참조하면, 본 개시의 각 실시예의 상술한 특징과 기타의 특징, 이점 및 측면은 더욱 분명해질 것이다. 첨부 도면에서, 동일 또는 유사한 도면 부호는 동일 또는 유사한 요소를 표시하고, 여기서:

도1은 전통적인 음성 인식 과정의 개략도를 도시한다.

도2는 본 개시의 실시예에 따른 예시적인 실시간 음성 인식 시나리오의 개략도를 도시한다.

도3은 본 개시의 실시예에 따른 절단 어텐션에 기반하는 실시간 음성 인식 방법의 흐름도를 도시한다.

도4는 본 개시의 실시예에 따른 절단 어텐션에 기반하는 실시간 음성 인식 과정의 개략도를 도시한다.

도5는 본 개시의 실시예에 따른 연결성 시계열 분류(CTC) 스파이크 정보에 기반하는 스트리밍 다층 절단 어텐션(streaming multi-layer truncated attention, SMLTA) 모델의 예시적인 아키텍처의 개략도를 도시한다.

도6은 본 개시의 실시예에 따른 CTC 스파이크 정보에 기반하는 SMLTA 모델의 예시적인 작동 원리도를 도시한다.

도7은 본 개시의 실시예에 따른 절단 어텐션에 기반하는 실시간 음성 인식 장치의 블록도를 도시한다.

도8은 본 개시의 복수의 실시예를 실시 가능한 전자 기기의 블록도를 도시한다.

발명을 실시하기 위한 구체적인 내용

[0014] 아래, 첨부 도면을 참조하여 본 개시의 실시예에 대해 더 상세하게 설명하고자 한다. 비록 첨부 도면에 본 개시의 어떤 실시예가 도시되었지만, 본 개시는 여러 가지 형식으로 구현 가능한 것으로, 여기서 서술하는 실시예에 한정되는 것으로 해석되어서는 안되고, 반대로, 이러한 실시예를 제공하는 것은 본 개시를 더욱 투철하게 그리고 완전하게 이해시키기 위한 것임을 이해하여야 한다. 본 개시의 첨부 도면 및 실시예는 예시적 작용을 위한 것으로, 본 개시의 보호 범위를 제한하기 위한 것이 아님을 이해하여야 한다.

[0015] 본 개시의 실시예의 설명에서, 전문 용어 '포함' 및 그 유사 용어는 개방적 포괄, 즉, '포함하나 이에 한정되지 않음'으로 이해하여야 한다. 전문 용어 '기반하여'는 '적어도 부분적으로 기반하여'로 이해하여야 한다. 전문 용어 '한 개 실시예' 또는 '당해 실시예'는 '적어도 한 개 실시예'로 이해하여야 한다. 전문 용어 '일부 실시예'는 '적어도 일부 실시예'로 이해하여야 한다. 후술은 또한, 기타의 명확한 및 묵시적인 정의를 포함할 수 있다.

[0016] 전통적인 음성 인식 시스템은 주로 음향 모델, 언어 모델 및 디코더 세 부분을 포함하는바, 도1에 도시한 바와 같이, 전통적인 음성 인식 과정(100)의 개략도가 도시된다. 수집된 음성 신호(110)가 획득된 후, 우선, 단계120에서 음성 신호(110)에 대하여 신호 처리 및 특징 추출을 수행하는바, 이에는 입력된 음성 신호(110)로부터 특징을 추출하여 음향 모델(132)이 이를 처리하도록 하는 단계가 포함되고, 일부 신호 처리 기술이 더 포함되어 환경 소음 또는 기타 요소가 특징에 주는 영향을 감소시킨다.

[0017] 도1을 참조하면, 특징 추출 단계 이후, 추출된 특징을 디코더(130)에 입력하고 디코더(130)에 의하여 처리를 거친 후 인식 결과(140)를 출력하고, 디코더(130)가 음향 모델(132) 및 언어 모델(134)에 따라 최대 확률로 출력되는 음성 신호의 단어 서열을 찾는바, 음향 모델(132)은 음성에서 음절로의 변환을 구현하고 언어 모델(134)은 음절에서 텍스트로의 변환을 구현한다. 음향 모델(132)은 음향, 음성학, 환경 특성 및 화자의 성별과 말씨 등 정보를 이용하여 음성 신호에 대하여 모델링한다. 예를 들면, 음향 모델(132)은 은닉 마르코프 모델 (HMM)을 적용하여 모델링함으로써 음성의 특징 벡터 서열의 어느 한 상태의 서열에 대한 사후확률을 표시할 수 있다. 언어 모델(134)은 언어에 대하여 모델링한다. 통상적으로는, 통계학의 N-그램 문법(N-Gram)을 사용할 수 있는바, 즉 전후 N개 단어가 나타나는 확률을 통계한다. 도1에 도시한 바와 같이, 음향 모델(132)은 음성 데이터베이스(133)에 기반하여 트레이닝되거나 및/또는 작동될 수 있고, 언어 모델(134)은 텍스트 데이터베이스(135)에 기반하여 트레이닝되거나 및/또는 작동될 수 있다.

[0018] 일반적으로, 모델의 복잡도와 계산량을 감소시키기 위하여, 음향 모델(132)과 언어 모델(134)은 통상적으로 2개의 독립적인 모델로 간주되어 각각 트레이닝 및 최적화된다. 도1에 도시한 바와 같이, 음성 인식 시스템의 구축은 특징 추출, 음향 모델 트레이닝, 언어 모델 트레이닝 및 디코더 최적화 등에 관한 복수의 모듈이 협동하여야 한다. 특히 음향 모델 트레이닝은 통상적으로 전문적인 음향 배경 지식을 갖춘 인원에 의하여 개발되고 최적화되어야 하는바, 따라서 음성 인식 시스템의 복잡도와 인건비가 증가하게 된다.

[0019] 신경망 기술이 부단히 발전됨에 따라, 음향 모델의 각 모듈이 점차 신경망에 의해 대체됨으로서, 음향 모델의 복잡도가 간략화될 뿐만 아니라 모델을 개발하고 디버깅하는 난이도가 낮아지고, 음성 인식 시스템의 성능이 현저히 향상된다. 예를 들면, 심층 신경망(DNN) + HMM식 음향 모델이 출현하였다. 이어서, 당해 구조에 기초하여 음향 모델에는 심층 합성곱망(CNN), 게이트 기반 순환(GRU) 신경망 및 장단기 메모리(LSTM) 망 등의 네트워크 구조가 더 도입되어 DNN 모델을 대체하게 되는데, 따라서 신경망 모델의 모델링 정확도가 현저히 향상되었다.

[0020] 연결성 시계열 분류(CTC) 모델은 엔드-투-엔드(end-to-end) 모델로, 대규모 단어량의 음성 인식에 사용되는바, 여기서는 DNN + HMM 혼합식 음향 모델 구조가 일체화된 신경망 구조로 완전히 대체됨으로써, 음향 모델의 구조 및 트레이닝 난이도가 대폭으로 간단화되어 음성 인식 시스템의 정확률이 더 향상된다.

[0021] 그러나, 본 출원의 발명자는 CTC 모델에 아래 몇 개 측면의 결함이 존재함을 인식하게 되었다. (1)비록 CTC 모델은 엔드-투-엔드 모델에 속하나 CTC 모델은 여전히 출력 서열 분포가 일정한 조건의 독립적 가설을 충족시키는 것이 요구된다. 이러한 독립적 가설로 인하여 CTC 모델은 음향 모델의 트레이닝 과정에서 효과적으로 언어 모델의 정보를 이용하여 모델링할 수 없게 되고; 입력된 음성 신호에 일정한 환경 소음이 존재하거나 또는 신호

세기가 보다 낮은 경우, 음성 인식 시스템은 단지 음향 정보에만 의존하여서는 입력된 오디오 특징에 대하여 효과적으로 구분하기가 매우 어려워져, 인식 성능이 저하되기 쉽다. (2)CTC 모델에 일정한 비율의 삽입 및 삭제 오류가 발생하기 쉽고, 이러한 오류는 비록 디코딩에 있어서 언어 모델의 가중치에 대하여 세심하게 설계함으로써 어느 정도 완화가 가능하지만, CTC 모델의 이러한 모델링 능력 부족 문제를 근원적으로 해결할 수는 없다.

[0022] 어텐션(attention) 모델은 인코더-디코더 모델에 대한 확장으로, 보다 긴 서열에서의 예측 효과를 향상시킬 수 있다. 우선 GRU 또는 LSTM 모델을 사용하여, 입력된 오디오 특징을 인코딩하여 잠재적 특징을 획득하고, 이어서 어텐션 모델에 의하여 이러한 잠재적 특징의 부동한 부분에 대하여 대응되는 가중치를 할당하고, 마지막으로 디코더에 의하여 부동한 모델링 입도에 따라 대응되는 텍스트를 출력한다. 이러한 음향 및 언어 모델의 합동 모델링 방식은 음성 인식 시스템의 복잡도를 더 간략화할 수 있다.

[0023] 그러나, 본 출원의 발명자는 어텐션 모델에는 여전히 아래 몇 개 측면의 결함이 존재함을 인식하게 되었다. (1)전통적인 어텐션 모델은 실시간 음성 인식을 구현할 수 없지만, 온라인의 음성 인터랙션 제품 서비스에서 음성은 하나 하나의 세그멘트가 사용자 기기로부터 서버에 업로드되는 것으로, 각 세그멘트는 도착되었을 때마다 당해 세그멘트에 대하여 디코딩이 요구되며, 이로써 디코딩 시간과 전송 과정이 병행될 수 있다. 이때 사용자가 말을 한 후 인식 결과를 대기하는 시간은 바로 마지막 패키지의 디코딩 시간이고, 이 과정은 스트리밍 디코딩으로 지칭된다('실시간 인식'으로 지칭되기도 함). 그러나 전통적인 어텐션 모델에서는 모든 음성 신호의 입력이 완료된 후에야 인식 시스템이 디코딩을 작동시킬 수 있는바, 사용자의 대기 시간이 바로 하나의 음성을 통째로 디코딩하는 시간이다. 따라서 전체 문구 어텐션 메커니즘에 기반하는 모델은 음성 인터랙션 시스템에서 사용자의 대기 시간을 보다 길어지게 하고 실시간 음성 인터랙션이 불가능하게 되며, 때문에 이가 음성 인터랙션 류의 산업 제품에 실제로 사용되는 것은 불가능하다. (2)전통적인 어텐션 모델은 긴 문구에 대한 모델링 효과가 양호하지 못한바, 인식 정확률이 저하되고, 전체 문구에 기반하는 어텐션 모델은 전체 문구 정보에 대하여 어텐션 모델링을 수행하여야 하나, 음성 인식의 문구 정보는 음성 프레임의 길이에 있어서는 가장 길어서 수천 프레임에 달할 수 있고 어텐션 모델링의 본질은 모델링 유닛에 대하여 중복 특징의 제거와 가장 핵심적인 본질적 특징의 선택을 수행하는 것이다. 만약 후보 특징의 범위가 지나치게 크고(예를 들어 전통적인 어텐션 모델은 수천 프레임의 범위 내에서 어텐션 모델링이 수행되어야 함) 또한 심층 학습의 수렴이 표준적인 컨벡스 최적화가 아니라면, 최종적으로 어텐션의 학습 과정은 당해 모델링 유닛을 제대로 설명 가능한 핵심 특성 프레임으로 수렴되기 어렵다. 따라서 어텐션 모델의 모델링 능력이 저하되고 긴 문구에 대한 모델링 효과가 양호하지 못하는 결함이 초래된다.

[0024] 이미 나타난 어텐션 모델에 대한 개선에는, 예를 들어, 입력 특징을 경험적(heuristic) 고정 길이에 따라 분할하는 것이 있다. 그러나, 이러한 고정 길이의 분할은 입력된 오디오 특징 서열 자체가 고려되지 않는바, 유효한 특징 서열을 배제시키기 쉽고, 음성 인식과 같이 말하는 속도와 멈춤이 부단히 변화되는 상황에 적응되기 어렵다. 따라서, 이러한 개선은 비록 부분적으로 실시간 음성 인식을 구현할 수 있지만 이는 음성 인식의 정확률을 희생시키는 것으로, 인식 정확률이 전통적인 어텐션 모델에 있어서의 전체 문구 모델링보다도 더 낮게 된다. 다시 말해, 전통적인 음성 인식 기술은 실시간 음성 인식이 불가능하거나 인식의 정확률이 보다 낮으므로 정확한 실시간 음성 인식을 구현할 수 없다.

[0025] 상기의 전통 기술에 존재하는 문제 중 하나 또는 복수를 해결하기 위하여, 본 개시의 실시예는 절단 어텐션에 기반하는 실시간 음성 인식 방안을 제공하는바, 이는 절단 정보를 이용하여 어텐션 모델 학습을 지도하는 구상을 제공한다. 본 개시의 실시예는 전통적인 어텐션 모델에 음성 신호에 따라 결정되는 절단 정보를 도입함으로써, 각 절단에 대하여 어텐션 모델링을 각각 수행하도록 어텐션 모델을 지도할 수 있는바, 연속적인 음성 인식을 구현할 수 있을 뿐만 아니라 높은 정확률을 보장할 수 있다. 또한, 선택적으로, 절단 정보는 CTC 처리에 의하여 획득된 음성 신호의 스파이크 정보에 기반할 수 있다. 또한, 선택적으로, 본 개시의 일부 실시예에 의하여 제공되는 CTC및 어텐션 듀얼 헤드 구조는 실시간 음성 인식 과정에서의 계산량이 보다 적도록 보장할 수 있고, 본 개시의 일부 실시예에 의하여 제공되는 복수 등급의 어텐션 구조는 음성 인식의 정확률을 더 향상시킬 수 있다. 아래 도2 내지 도8을 참조하여 본 개시의 일부 예시적인 구현에 대해 상세히 설명하고자 한다.

[0026] 도2는 본 개시의 실시예에 따른 실시간 음성 인식 시나리오(200)의 개략도를 도시한다. 도시된 시나리오(200)는 본 개시의 실시예를 구현 가능한 하나의 예시적인 시나리오일 뿐, 본 개시의 보호 범위를 한정하고자 하는 것이 아님을 이해하여야 한다.

[0027] 도2에 도시한 바와 같이, 시나리오(200)에서, 사용자(210)는 그의 사용자 기기(220)에 마주하여 말하고 있고, 사용자(210)에 의하여 발생하는 음성(215)(즉, 소리)은 사용자 기기(220)에 의하여 수집되는바, 예를 들면, 사

용자 기기(220)의 소리 수집 기기(예를 들면 마이크로폰)에 의하여 음성(215)이 수집될 수 있다. 사용자 기기(220)는 음성 신호를 수집할 수 있는 임의의 전자 기기일 수 있는바, 스마트 폰, 태블릿 컴퓨터, 데스크톱 컴퓨터, 랩톱 컴퓨터, 스마트 웨어러블 기기(예를 들면 스마트 손목시계, 스마트 안경), 네비게이션 기기, 멀티미디어 플레이어 기기, 교육 기기, 게임 기기, 스마트 스피커 등을 포함하나 이에 한정되지 않는다.

[0028] 사용자 기기(220)는 음성(215)을 수집하는 과정에서, 네트워크(225)를 통해 음성(215)을 세그먼트로 나누어 서버(230)로 송신할 수 있는바, 서버(230)는 본 개시의 실시예에 따른 음성 인식에 사용되는 절단 어텐션 모델(233)(본 개시의 일부 실시예에서, 절단 어텐션 모델(233)은 CTC 스파이크 정보에 기반하는 스트리밍 다층 절단 어텐션 모델일 수 있음)을 포함한다. 네트워크(225)는 임의의 유선 네트워크, 무선 네트워크, 또는 유선 네트워크와 무선 네트워크의 조합일 수 있다. 절단 어텐션 모델(233)은 트레이닝 데이터(235)에 의하여 트레이닝되어 실시간으로 정확한 음성 인식을 구현할 수 있고, 인식이 완료된 후, 네트워크(225)를 통해 인식 결과를 사용자 기기(220)로 송신할 수 있다.

[0029] 사용자 기기(220)는 디스플레이(223)를 통해 인식 결과를 실시간으로 표시할 수 있다. 본 개시의 실시예는 음성을 실시간으로 인식할 수 있으므로 음성(215)이 부단히 발생됨에 따라, 디스플레이(223)에 표시되는 인식 결과도 동적으로 변화되는바, 사용자로 하여금 인식되는 결과을 실시간으로 인지할 수 있도록 하여 사용자 경험(user experience)을 향상시킨다. 일부 실시예에서, 사용자 기기(220)는 디스플레이(223)를 포함하지 않을 수도 있는바, 반대로, 사용자 기기(220)는 이의 오디오 출력 기기(예를 들면, 스피커)를 통해 인식 결과를 출력할 수 있다. 다른 일부 실시예에서, 서버(230)는 음성에 대한 인식 결과를 사용자 기기(220)로 송신할 필요 없이 인식 결과를 기반으로 더 많은 동작을 수행할 수 있다.

[0030] 도2의 시나리오(200)에서는 절단 어텐션 모델(233)이 사용자 기기(220)의 원격 서버(230)측에 배치되는 것으로 도시되었지만, 절단 어텐션 모델(233)은 사용자 기기(220)의 로컬에 배치될 수도 있음을 이해하여야 한다. 대안적으로, 절단 어텐션 모델(233)의 일부분을 사용자 기기(220) 로컬에 배치하고 다른 부분을 서버(230) 측에 배치할 수도 있고, 또는 절단 어텐션 모델(233)을 분산식으로 복수의 서버에 배치할 수도 있다. 본 개시의 실시예는 절단 어텐션 모델(233)의 배치 방식 및 위치에 대하여 한정하지 않는다.

[0031] 도3은 본 개시의 실시예에 따른 절단 어텐션에 기반하는 실시간 음성 인식 방법(300)의 흐름도를 도시한다. 방법(300)은 상기의 도2를 참조하여 설명한 서버(230), 사용자 기기(220) 또는 이들의 조합에 의하여 수행될 수 있음을 이해하여야 한다. 방법(300)을 명확히 설명하는데 편의를 위하여, 도4의 절단 어텐션에 기반하는 실시간 음성 인식 과정(400)을 참조하여 함께 방법(300)을 설명하고자 한다.

[0032] 단계302, 입력된 음성 신호에 기반하여, 음성 신호의 특징 서열을 절단하는 절단 정보를 획득한다. 예를 들어, 특징 서열은 음성 신호를 인코딩하여 획득할 수 있다. 예를 들면, 도4에 도시한 바와 같이, 사용자로부터의 음성 신호(410)가 수집된 후, 인코더(420)는 수집된 음성 신호(410)를 잠재적 특징 서열(430)로 인코딩하고, 잠재적 특징 서열(430)에 따라 절단 정보(435)를 결정할 수 있다. 본 개시의 실시예는 입력된 음성 신호의 특징에 따라 산출되는 절단 정보를 사용함으로써, 후속의 절단이 음성 인식의 정확률에 영향 주지 않도록 하므로 음성 신호에 대한 실시간 인식이 구현되면서 음성 인식의 정확률도 보장된다.

[0033] 일부 실시예에서, 정보 절단은 특징 서열에 대하여 연결성 시계열 분류(CTC) 처리를 수행하여 음성 신호에 관련되는 스파이크 정보를 획득하는 것일 수 있는바, CTC 처리는 스파이크의 서열을 출력할 수 있고 각 스파이크 사이는 블랭크(blank)로 구획될 수 있으며, 하나의 스파이크는 하나의 음절(syllable) 또는 한 그룹의 음소(phoneme), 예를 들어, 고주파 음소의 조합을 표시할 수 있다. 비록 본 명세서에서는 아래에 CTC 스파이크 정보를 절단 정보를 제공하는 하나의 예시로 사용하여 설명하지만, 현재 이미 알려진 또는 앞으로 개발될, 입력된 음성 신호의 절단 정보를 제공할 수 있는 임의의 기타 모델 및/또는 알고리즘도 본 개시의 실시예에 결합되어 사용될 수 있음을 이해하여야 한다.

[0034] 단계304에서, 절단 정보에 기반하여 특징 서열을 복수의 서브서열로 절단한다. 예를 들면, 도4에 도시한 바와 같이, 어텐션에 기반하는 디코더(440)는 절단 정보(435)에 따라 잠재적 특징 서열(430)을 하나 하나의 잠재적 특징 서브서열로 절단할 수 있는바, 여기서 잠재적 특징 서열은 음성 신호의 특징을 표시하는 벡터일 수 있다. 예를 들면, 잠재적 특징 서열은 직접 관측으로 획득될 수는 없으나 관측 가능 변수에 의하여 결정 가능한 특징 벡터를 지칭할 수 있다. 전통적인 기술에서 고정 길이의 절단 방식을 사용하는 것과는 달리, 본 개시의 실시예는 음성 신호에 의하여 결정되는 절단 정보를 사용하여 특징 절단을 수행하는바, 이는 유효한 특징 부분이 배제되는 것을 회피할 수 있고 따라서 매우 높은 정확률을 보장할 수 있다.

[0035] 단계306에서, 복수의 서브서열 중의 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간 인식 결과를 획득한다. 예를 들면, 디코더(440)는 절단으로 생성되는 각 잠재적 특징 서브서열에 대하여, 어텐션 모델을 사용하여 인식 결과(450)를 획득하는바, 여기서 어텐션 모델은 가중 특징 선택을 구현하고 잠재적 특징의 부동한 부분에 대하여 대응되는 가중치를 할당할 수 있다. 현재 이미 알려진 또는 앞으로 개발될, 어텐션 메커니즘에 기반하는 임의의 모델 및/또는 알고리즘은 본 개시의 실시예에 결합되어 사용될 수 있다. 따라서, 본 개시의 실시예는 전통적인 어텐션 모델에 음성 신호에 따라 결정되는 절단 정보를 도입함으로써, 각 절단에 대하여 어텐션 모델링을 각각 수행하도록 어텐션 모델을 지도할 수 있는바, 연속적인 음성 인식을 구현할 수 있을 뿐만 아니라 높은 정확률을 보장할 수 있다.

[0036] 일부 실시예에서는, 잠재적 특징 서열이 복수의 서브서열로 절단되는 단계 이후, 복수의 서브서열 중의 제1 서브서열에 대하여 어텐션 모델의 제1 어텐션 모델링을 수행하고, 복수의 서브서열 중의 제2 서브서열에 대하여 어텐션 모델의 제2 어텐션 모델링을 수행할 수 있는바, 제1 어텐션 모델링과 제2 어텐션 모델링은 상이하다. 다시 말해, 본 개시의 실시예는 국소적으로 절단되는 어텐션 모델의 어텐션 모델링을 수행할 수 있다.

[0037] 도5는 본 개시의 실시예에 따른 CTC 스파이크 정보에 기반하는 SMLTA 모델의 아키텍처(500)의 개략도를 도시한다. 도5에 도시한 바와 같이, 아키텍처(500)는 주로, 입력된 음성 신호(510)를 잠재적 특징 서열(530)(즉, 잠재적 특징 서열(h))로 인코딩하도록 구성되는 공유 인코더(520); 잠재적 특징 서열(530)에 기반하여 스파이크 정보(560)를 결정하도록 구성되는 CTC 모듈(540); 및 잠재적 특징 서열(530) 및 스파이크 정보(560)에 기반하여 실시간 인식 결과(570)를 획득하도록 구성되는 어텐션 디코더(550); 이 세 부분을 포함한다. 도5에 도시한 아키텍처(500)에서, 공유 인코더(520)는 CTC 모듈(540) 및 어텐션 디코더(550)에 공유되고 CTC 모듈(540) 및 어텐션 디코더(550)의 2개의 헤드인 '듀얼 헤드 구조'를 형성한다. 아키텍처(500)에서 주요한 계산량은 공유 인코더(520)와 어텐션 디코더(550)에 집중되므로 전통적인 어텐션 모델에 비하면, 본 개시의 실시예에 따른 아키텍처(500)는 전체 모델 아키텍처의 계산량을 거의 증가시키지 않으면서(증가가 있긴 하나, 증가된 비율이 매우 작은바, 이는 거의 무시 가능함) CTC 출력(즉, 스파이크 정보(560))과 어텐션 출력(즉, 인식 결과(570))을 동시에 제공 가능하고 본 개시의 실시예의 아키텍처(500)의 대규모 산업 배치 문제를 잘 해결할 수 있다. 반대로, 전통적인 방식에서의 제품에 이러한 두 모델을 조합하여 사용하려면 온라인 계산량이 배로 증가될 수 있으며 따라서 비용이 지나치게 높아지고 대규모의 산업적 사용이 불가능하다.

[0038] 도5에 도시한 바와 같이, 공유 인코더(520)는 하나의 합성곱층(Conv), N개의 LSTM과 배치 정규화(BN)층을 포함하는바, 여기서 N은 양의 정수(예를 들어, 5 등)일 수 있고, LSTM은 일방향의 LSTM일 수 있다. 주어지는 입력 음성 신호에 대하여 공유 인코더(520)는 우선 음성 신호를 인코딩하여 대응되는 잠재적 특징 서열(530)을 획득한다. 일부 실시예에서, 음성 신호(510)는 공유 인코더(520)에 입력되기 전에 이미 특징 추출 처리되어 모델 입력(x)으로 사용된다. 비록 도5에서 공유 인코더(520)의 내부 계층 구조를 도시하였지만 기타 구조의 인코더도 본 개시의 실시예에 결합되어 사용될 수 있음을 이해하여야 한다.

[0039] CTC 모듈(540)은 하나의 선형 변환(Linear)층과 하나의 정규화(Softmax)층을 포함하는바, 이는 CTC 트레이닝 기준을 이용하여 입력된 잠재적 특징 서열(530)의 스파이크에 대한 설명 정보를 획득함으로써 스파이크 정보(560)를 포함하는 CTC 출력을 생성하고, 스파이크 정보(560)를 어텐션 디코더(550)에 전달하여 스파이크 정보(560)가 잠재적 특징 서열(530)을 복수의 서브서열로 절단하는데 이용되도록 한다.

[0040] 계속하여 도5를 참조하면, 어텐션 디코더(550)는 하나의 어텐션 모델층, M개의 LSTM층과 층 정규화(LN)층 및 하나의 Softmax층을 포함하는바, 여기서 M은 양의 정수(예를 들어, 2 등)일 수 있고 LSTM은 일방향의 LSTM일 수 있다. 어텐션 디코더(550)는 수신되는 잠재적 특징 서열(530) 및 스파이크 정보(560)에 따라 잠재적 특징 서열(530)을 하나 하나의 연속되는 서브서열로 절단할 수 있다. 어텐션 디코더(550)는 절단된 서브서열에 대하여 어텐션 메커니즘으로 선별을 수행하여 최종적으로 대응되는 출력 확률 분포를 획득한다. 공유 인코더(520)와 디코더(550)는 일방향 LSTM을 기본적인 네트워크 구조로 사용하고 있고 또한 이러한 잠재적 특징 서열을 서브서열로 절단하는 방식은 단지 잠재적 특징의 역사적 정보에 의존하는 것이다. 따라서, 이는 음성 인식 시스템으로 하여금 오디오 신호가 입력되는 동시에 실시간으로 디코딩할 수 있도록 하여, 전체 오디오 신호가 모두 입력되고 난 후에야 디코딩을 시작하지 않아도 되고, 실시간 음성 인식을 구현한다. 또한, 도5에 도시된 어텐션 디코더(550)의 내부 계층 구조외에 기타 구조의 어텐션 디코더도 본 개시의 실시예에 결합되어 사용될 수 있음을 이해하여야 한다.

[0041] 도5에 도시한 콤팩트한 CTC와 어텐션 일체형 STMLA 모델은, CTC 모듈(540)의 헤드와 어텐션 디코더(550)의 헤드를 동시에 보유하므로, 트레이닝하는 경우, 듀얼 헤드로 함께 트레이닝할 수도 있다. CTC 모듈(540)의 헤드는

CTC손실 함수로 트레이닝하고, 어텐션 디코더(550)의 헤드는 크로스 엔트로피 손실 함수로 트레이닝한다. 두 손실 함수는 선형 보간의 가중치로 연결되는바, 즉, 두 손실 함수는 각각 그에 대응되는 가중치를 가진다. 이러한 모델은 온라인 계산량이 하나의 모델과 거의 같으나 CTC와 어텐션 두 가지 정보를 동시에 제공 가능하고 따라서 대규모 산업 배치가 가능하다.

[0042] CTC 모듈에 의하여 출력되는 스파이크 정보에는 일정한 삽입 및 삭제 오류가 존재하는바, 이러한 오류로 인하여, 절단에 의해 획득된 잠재적 특징 서열의 서브서열 경계가 부정확하게 되기 쉬우며, 따라서 음성 인식 시스템의 인식 성능이 영향받게 된다. CTC 모듈의 삽입 및 삭제 오류로 인한 편차를 완화시키기 위하여, 본 개시의 실시예는 복수 등급의 어텐션 구조에 기반하여 절단되는 서브서열에 대하여 단계적으로 선별하는 방법을 더 제공하는바, 이는 어텐션 디코더의 LSTM 모델을 통해 최후의 확률 분포를 출력한다. 예를 들면, 도6은 본 개시의 실시예에 따른 2개 등급의 어텐션 구조에 기반하는 SMLTA 모델의 작동 원리도(600)를 도시한다.

[0043] 도6에 도시한 바와 같이, CTC의 스파이크 정보(610)를 이용하여 잠재적 특징 서열(620)을 복수의 특징 서브서열로 절단할 수 있는바, 여기서 *trunc*는 절단되는 기본 단위이고 하나의 *trunc*는 잠재적 특징 서열(620)에서 하나의 스파이크에 대응되는 특징 부분일 수 있다. 일부 실시예에서는, 각 스파이크(이는 하나의 음절 또는 고주파 음소의 조합을 표시할 수 있음)에 대하여 당해 스파이크와 인접하는 소정 개수의 스파이크에 대응되는 특정 서브서열을 하나의 절단으로 선택할 수 있는바, 소정 개수의 스파이크는 당해 스파이크 이전의 제1 개수의 스파이크와 당해 스파이크 이후의 제2 개수의 스파이크를 포함한다. 서브서열은 이러한 스파이크에 대응되는 특정 부분일 수 있다. 통상적으로는, 어느 스파이크에 있어서 시계열적으로 그 전후로 충분한 스파이크가 존재한다면, 제1 개수와 제2 개수의 총합은 소정 개수와 같다. 물론, 당해 스파이크 이전에 제1 개수의 스파이크가 존재하지 않는다면 당해 스파이크 이전의 모든 스파이크를 선택하고 마찬가지로 당해 스파이크 이후에 제2 개수의 스파이크가 존재하지 않는다면 당해 스파이크 이후의 모든 스파이크를 선택한다.

[0044] 화살표(630)에 도시한 바와 같이, 2개 등급의 어텐션 모델링을 수행한다. 예를 들면, i번째 스파이크에 대하여 잠재적 특징 서열(620) 중의 복수의 중첩되는 서브서열(예를 들면, 도6의 예시에서는 3개의 서브서열인바, 각각 trunc1에서 trunc2까지, trunc1에서 trunc3까지 및 trunc1에서 trunc4까지임)에 기반하여 제1 등급 어텐션 메커니즘(예를 들면, 도6의 attention1)을 통해 복수의 부동한 특징 벡터(도6의 예시에서는 각각 $(c_{i,1})$, $(c_{i,2})$, $(c_{i,3})$인바 이러한 특징 벡터는 각각 대응되는 어텐션 벡터인 $(\alpha_{i,1})$, $(\alpha_{i,2})$, $(\alpha_{i,3})$을 구비함)를 각각 결정할 수 있다. 다음, 결정된 복수의 특징 벡터(즉, $(c_{i,1})$, $(c_{i,2})$, $(c_{i,3})$)를 기반으로 제2 등급 어텐션 메커니즘(예를 들면, 도6의 attention2)에 의하여 나아가 문맥 특징 벡터(c_i는 대응되는 어텐션 벡터 β_i를 구비함)를 선별 및 결정한다. 비록 도6에서 3개의 특징 벡터 $(c_{i,1})$, $(c_{i,2})$, $(c_{i,3})$을 발생시키는 것에 대하여 도시하였으나 기타 개수의 특징 벡터도 사용 가능한 바, 이는 정확률과 하드웨어 디코딩 지연 시간에 따라 결정될 수 있는 것으로, 본 개시의 실시예는 백터 개수에 한정되지 않음을 이해하여야 한다.

[0045] 계속하여 도6을 참조하면, 이어서, 디코더의 디코딩 과정이 수행되는바, 화살표(640)에 도시한 바와 같이, 문맥 특징 벡터(c_i)를 디코더(예를 들어, RNN, LSTM 또는 GRU 등)에 입력하여 인식 결과(y_i)를 출력하며, 여기서 도6의 s_i는 출력 상태를 표시하고 h'_i는 잠재 상태를 표시한다. 이러한 방식으로, 2개 등급의 어텐션 구조를 통하여 CTC의 삽입 및 삭제 문제를 효과적으로 해결하고 SMLTA 모델의 음성 인식 성능을 효과적으로 향상시킬 수 있다.

[0046] 본 개시의 실시예에 따르면, CTC 스파이크 정보를 이용하여 어텐션 모델에 절단 정보를 제공하는 모델링 방법은, 전통적인 어텐션 모델이 긴 문구에 대한 모델링 정확도가 양호하지 못한 문제를 해결하고, 어텐션 모델링 기술을 실시간 음성 인식의 스트리밍 디코딩에 적용하려는 산업 분야의 요구를 최초로 충족시키며, 세계적인 범위에서 산업 분야 최초로 높은 정확도의 어텐션 모델을 스트리밍 음성 서비스에 적용하는 것을 구현한다. 현재 가장 우수한 CTC 모델 기반의 음성 인식 시스템에 비하면, 본 개시의 일부 실시예에 의하여 제공되는 CTC 스파이크 정보에 기반하는 SMLTA 모델은 인식 오류율이 15%이상 감소된다.

[0047] 따라서, 본 개시의 일부 실시예에 따르면, 절단에 기반하는 엔드-투-엔드 실시간 음성 인식 방법이 제공되는바, 이는 문맥과 무관한 음절을 기본적인 음향 모델링 유닛으로 하고 CTC와 어텐션 합동 모델링 방식으로, 우선 CTC의 스파이크 정보를 이용하여 잠재적 특징 서열이 절단되는 위치 정보를 획득한 다음, 어텐션 모델링 기술을 도입하여 엔드-투-엔드 음성 인식 방법을 구현한다. 당해 방법은 특정 위치에 절단 정보를 도입함으로써, 전통적인 어텐션 전체 문구 모델링 방법에 있어서 실시간 디코딩을 구현할 수 없는 결함을 회피할 수 있을 뿐만 아니라 어텐션 엔드-투-엔드 모델링에 의하여 더 높은 모델 정확도 우세를 구비하게 되는바, 음성 인식 제품의 성능

을 대폭 향상시킨다. 이러한 절단에 기반하는 엔드-투-엔드 음성 인식 모델링 방법은 대규모 단어량의 실시간 연속적인 음성 인식에 엔드-투-엔드 모델링 기술을 적용하는 구조적인 창조를 최초로 구현한다.

[0048] 또한, 본 개시의 일부 실시예는 CTC 스파이크 정보에 기반하는 스트리밍 다층 절단 어텐션(SMLTA) 모델을 더 제공한다. 당해 모델은 문맥과 무관한 음절 또는 고주파 음소의 조합을 기본적인 음향 모델링 유닛으로 하고, CTC와 어텐션 모델의 합동 모델링 기술에 의하여, 예를 들면, 수만 시간의 트레이닝 데이터의 조건 하에 CTC 모델을 현저히 초월하는 음성 인식 성능을 가진다. 전통적인 어텐션 모델에 있어서 전체 문구를 모두 인코딩한 잠재적 특징이 획득된 후에야 각 모델링 유닛의 어텐션 모델을 모델링할 수 있는 것에 비해, 본 개시의 실시예는, CTC 모델의 출력 결과(스파이크 정보로도 지칭됨)를 이용하여, 연속적으로 인코딩되는 잠재적 특징 서열을 하나 하나의 잠재적 특징 서브서열로 절단하는바, 각 모델링 유닛의 어텐션 모델의 모델링은 오직 하나의 독립적인 잠재적 특징 서브서열에서 독립적으로 진행되는 것을 창조적으로 제기하였다. 이러한 모델은 인코딩되는 잠재적 특징에 대하여 전체 문구 길이 범위 내에서 어텐션 모델링할 필요가 없고, 어텐션 모델링 과정이 하나 하나의 작은 특징 서브서열로 제약되어 진행되므로, 전통적인 어텐션 모델에 있어서 큰 범위 내에서 어텐션 모델링하는 경우, 정확도가 양호하지 못한 문제를 성공적으로 해소한다. 또한, CTC 모델이 출력하는 스파이크 정보는 왼쪽에서 오른쪽으로의 순서로 스트리밍 전개되는 것이므로 상술한 서브서열도 왼쪽에서 오른쪽으로의 순서로 스트리밍 전개되고, 따라서 최종적으로 본 개시의 SMLTA 모델도 왼쪽에서 오른쪽으로의 순서로 스트리밍 진행될 수 있다.

[0049] 따라서, 전통적인 음성 인식 기술 또는 모델에 비하면, 본 개시의 실시예는 다음과 같은 이점이 있다.

[0050] 전통적인 CTC 모델에 비하면, 본 개시의 실시예는 CTC와 어텐션 합동 트레이닝의 방법에 의하여 음향 모델과 언어 모델의 합동 모델링을 효과적으로 구현하는바, 전통적인 CTC 모델이 음향 모델에 대한 모델링만 가능한 결함을 해소하고 복잡한 환경에서의 음성 인식 시스템의 강건성을 향상시킨다. 또한, CTC 모델에는 삽입 및 삭제의 문제가 존재하므로 CTC 모델이 어텐션 모델에 제공하는 정보에는 오류가 존재할 수도 있다. 그러므로 본 개시의 일부 실시예는 2개 등급의 어텐션 구조로 CTC 모델의 결함을 효과적으로 해결하고 음성 인식의 성능을 현저히 향상시킨다.

[0051] 전통적인 어텐션 모델에 비하면, 본 개시의 일부 실시예는 음성 신호의 절단 정보를 이용하여 절단을 수행하는 어텐션 모델을 이용하는바, 전통적인 어텐션 모델로 구축되는 음성 인식 시스템에 있어서 실시간 디코딩이 불가능한 문제를 해결하고, 시스템 응답 속도에 대한 사용자 수요를 충족시킨다. 또한, 전통적인 어텐션 모델은 긴 문구에 대한 인식 정확률이 보다 낮은 것에 비해, 본 개시의 실시예는 절단 정보를 이용하여 각 모델링 유닛의 어텐션 모델링 과정을 각각의 특징 서브서열로 집중하여 모델링할 수 있으므로, 어텐션 모델의 모델링 정확도를 향상시키고 큰 범위 내에서 어텐션 모델 모델링 정확도가 양호하지 못한 문제를 해결한다.

[0052] 전통적인 고정 길이 분할을 지원하는 어텐션 모델에 비하면, 본 개시의 실시예는 정보 절단을 통해 특징 서브서열을 획득하는바, 이는 고정 길이 분할보다 더욱 정확하게 필요한 특징 서브서열을 획득할 수 있다. 또한, 본 개시의 일부 실시예는 2개 등급의 어텐션으로 특징 서브서열의 중첩을 구현하는바, 따라서 더 효과적으로 특징을 선별할 수 있고 모델링 정확도가 전체 문구 어텐션 방법을 초월하는 성능을 얻을 수 있다.

[0053] 또한, 본 개시의 일부 실시예에 의하여 제공되는 가장 콤팩트한 CTC와 어텐션 일체형 절단 어텐션 모델은 CTC의 헤드와 어텐션의 헤드를 동시에 보유하고 트레이닝하는 경우에도 듀얼 헤드로 함께 트레이닝한다. CTC의 헤드는 CTC의 손실 함수로 트레이닝하고, 어텐션의 헤드는 크로스 엔트로피 손실 함수로 트레이닝하는바, 두 손실 함수는 선형 보간의 가중치로 연결된다. 이러한 모델은 온라인 계산량이 하나의 모델과 거의 같으므로 추가적인 계산 비용 증가를 회피하고 따라서 대규모 산업 배치의 필요를 충족시킨다.

[0054] 도7은 본 개시의 실시예에 따른 절단 어텐션에 기반하는 실시간 음성 인식 장치(700)의 블록도를 도시한다. 도7에 도시한 바와 같이, 장치(700)는 절단 정보 획득 모듈(710), 특징 서열 절단 모듈(720) 및 인식 결과 획득 모듈(730)을 포함한다. 절단 정보 획득 모듈(710)은 입력된 음성 신호에 기반하여 음성 신호의 특징 서열을 절단하는 절단 정보를 획득하도록 구성된다. 특징 서열 절단 모듈(720)은 절단 정보에 기반하여 특징 서열을 복수의 서브서열로 절단하도록 구성된다. 인식 결과 획득 모듈(730)은 복수의 서브서열 중의 각 서브서열에 대하여 어텐션 메커니즘을 통해 실시간 인식 결과를 획득하도록 구성된다.

[0055] 일부 실시예에서, 절단 정보 획득 모듈(710)은, 특징 서열에 대하여 연결성 시계열 분류(CTC) 처리를 수행하여, 음성 신호에 관련되는 스파이크 정보를 획득하도록 구성되는 스파이크 정보 획득 모듈; 및 획득된 스파이크 정보에 기반하여 절단 정보를 결정하도록 구성되는 절단 정보 결정 모듈; 을 포함한다.

[0056] 일부 실시예에서, 인식 결과 획득 모듈(730)은, 복수의 서브서열 중의 제1 서브서열에 대하여 어텐션 모델의 제1 어텐션 모델링을 수행하도록 구성되는 제1 어텐션 모델링 장치 - 어텐션 모델은 가중 특징 선택을 구현 가능함 - ; 및 복수의 서브서열 중의 제2 서브서열에 대하여 어텐션 모델의 제2 어텐션 모델링을 수행하도록 구성되는 제2 어텐션 모델링 장치 - 제1 어텐션 모델링과 제2 어텐션 모델링은 상이함 - ; 를 포함한다.

[0057] 일부 실시예에서, 장치(700)는 음성 신호를 특징 서열로 인코딩하도록 구성되는 공유 인코더(520); 특징 서열에 기반하여 스파이크 정보를 획득하도록 구성되는 연결성 시계열 분류(CTC) 모듈; 및 특징 서열 및 스파이크 정보에 기반하여 실시간 인식 결과를 획득하도록 구성되는 어텐션 디코더(550); 를 포함하는바, 공유 인코더(520)는 CTC 모듈 및 어텐션 디코더(550)에 공유되고, 절단 정보 획득 모듈(710)은 CTC 모듈에 포함되고, 특징 서열 절단 모듈(720)과 인식 결과 획득 모듈(730)은 어텐션 디코더(550)에 포함된다.

[0058] 일부 실시예에서, 장치(700)는, CTC 모듈을 트레이닝하는 제1 손실 함수를 결정하도록 구성되는 제1 손실 함수 결정 모듈; 어텐션 디코더를 트레이닝하는 제2 손실 함수를 결정하도록 구성되는 제2 손실 함수 결정 모듈; 및 제1 손실 함수와 제2 손실 함수를 사용하여 CTC 모듈과 어텐션 디코더(550)를 함께 트레이닝하도록 구성되는 트레이닝 모듈; 을 더 포함한다.

[0059] 일부 실시예에서, 특징 서열 절단 모듈(720)은, 스파이크 정보 중의 각 스파이크에 대하여 특징 서열에서 각 스파이크와 인접하는 소정 개수의 스파이크에 대응되는 서브서열을 선택하도록 구성되는 서브서열 선택 모듈을 포함하고, 소정 개수의 스파이크는 각 스파이크 이전의 제1 개수의 스파이크와 각 스파이크 이후의 제2 개수의 스파이크를 포함한다.

[0060] 일부 실시예에서, 인식 결과 획득 모듈(730)은, 특징 서열 중의 복수의 중첩되는 서브서열에 기반하여 제1 등급 어텐션 메커니즘을 통해 복수의 특징 벡터를 각각 결정하도록 구성되는 복수의 특징 벡터 결정 모듈; 결정된 복수의 특징 벡터를 기반으로 제2 등급 어텐션 메커니즘에 의하여 문맥 특징 벡터를 결정하도록 구성되는 문맥 특징 벡터 결정 모듈; 및 결정된 문맥 특징 벡터에 기반하여 실시간 인식 결과를 결정하도록 구성되는 인식 결과 결정 모듈; 을 포함한다.

[0061] 일부 실시예에서, 특징 서열은 잠재적 특징 서열이고, 각 스파이크는 하나의 음절 또는 한 그룹의 음소를 나타낸다.

[0062] 도7에 도시한 절단 정보 획득 모듈(710), 특징 서열 절단 모듈(720) 및 인식 결과 획득 모듈(730)은 하나의 또는 복수의 전자 기기에 포함될 수 있음을 이해하여야 한다. 또한, 도7에 도시한 모듈은 본 개시의 실시예를 참조한 방법 또는 과정 중의 단계 또는 동작을 수행할 수 있음을 이해하여야 한다.

[0063] 도8은 본 개시의 실시예를 구현하기 위한 예시 기기(800)의 개략적 블록도를 도시한다. 기기(800)는 본 개시에서 설명하는 절단 어텐션에 기반하는 실시간 음성 인식을 구현하는 장치(700) 사용자 기기(220) 또는 서버(230)일 수 있음을 이해하여야 한다. 도시한 바와 같이, 기기(800)는 중앙 처리 유닛(CPU)(801)을 포함하는바, 이는 읽기 전용 메모리(ROM)(802)에 저장된 컴퓨터 프로그램 명령어 또는 저장 유닛(808)으로부터 랜덤 액세스 메모리(RAM)(803)로 로딩된 컴퓨터 프로그램 명령어에 따라 여러 가지 적당한 동작과 처리를 실행할 수 있다. RAM(803)에는 기기(800) 조작에 필요한 여러 가지 프로그램과 데이터가 저장될 수도 있다. CPU(801), ROM(802) 및 RAM(803)는 버스(804)를 통하여 서로 연결된다. 입력/출력(I/O) 인터페이스(805)도 버스(804)에 연결된다.

[0064] 기기(800) 중의 복수의 부품은 I/O 인터페이스(805)에 연결되는바, 입력 유닛(806), 예를 들어 키보드, 마우스 등; 출력 유닛(807), 예를 들어 여러 가지 유형의 디스플레이, 스피커 등; 저장 유닛(808), 예를 들어 자기 디스크, 광디스크 등; 및 통신 유닛(809), 예를 들어 네트워크 카드, 모뎀, 무선 통신 송수신기 등이 포함된다. 통신 유닛(809)은 기기(800)가 인터넷과 같은 컴퓨터 네트워크 및/또는 여러 가지 전기 통신 네트워크를 통하여 기타 기기와 정보/데이터를 교환하도록 허용한다.

[0065] CPU(801)는 전술에서 설명된 각각의 방법과 과정을 실행하는바, 예를 들어 방법(300)이다. 예를 들어, 일부 실시예에서, 방법은 컴퓨터 소프트웨어 프로그램으로 구현될 수 있는바, 이는 기계 판독 가능 매체, 예를 들어 저장 유닛(808)에 유형적으로 포함된다. 일부 실시예에서, 컴퓨터 프로그램의 일부 또는 전부가 ROM(802) 및/또는 통신 유닛(809)을 통하여 기기(800)에 로딩 및/또는 설치될 수 있다. 컴퓨터 프로그램이 RAM(803)에 로딩되고 CPU(801)에 의하여 실행될 시, 전술에서 설명된 방법의 하나 또는 복수의 동작 또는 단계를 실행할 수 있다. 대안적으로, 기타 실시예에서, CPU(801)는 기타 임의의 적당한 방식을 통해(예를 들어, 펌웨어의 도움으로) 방법을 실행하도록 구성될 수 있다.

[0066] 본 고안에서, 이상에서 설명된 기능은 적어도 부분적으로 하나 또는 복수의 하드웨어 논리 부품에 의해 실행될 수 있다. 예를 들어, 무제한적으로, 사용 가능한 시범 유형의 하드웨어 논리 부품은 필드 프로그램 가능 게이트 어레이(FPGA), 주문형 반도체(ASIC), 특정 용도 표준 제품(ASSP), 시스템 온 칩(SOC), 복합 프로그래머블 논리 소자(CPLD), 등을 포함한다.

[0067] 본 개시의 방법을 구현하기 위한 프로그램 코드는 하나 또는 복수의 프로그래밍 언어의 임의의 조합을 적용하여 작성할 수 있다. 이러한 프로그램 코드는 범용 컴퓨터, 전용 컴퓨터 또는 기타 프로그램 가능 데이터 처리 장치의 프로세서 또는 제어기에 제공되어 프로그램 코드가 프로세서 또는 제어기에 의해 실행될 시 흐름도 및/또는 블록도에 규정한 기능/조작이 구현되도록 할 수 있다. 프로그램 코드는 기계에서 전부 실행되거나, 기계에서 일부 실행되거나, 독립적인 소프트웨어 패키지로서 일부는 기계에서 실행되고 일부는 원격 기계에서 실행되거나, 혹은 원격 기계 또는 서버에서 전부 실행될 수 있다.

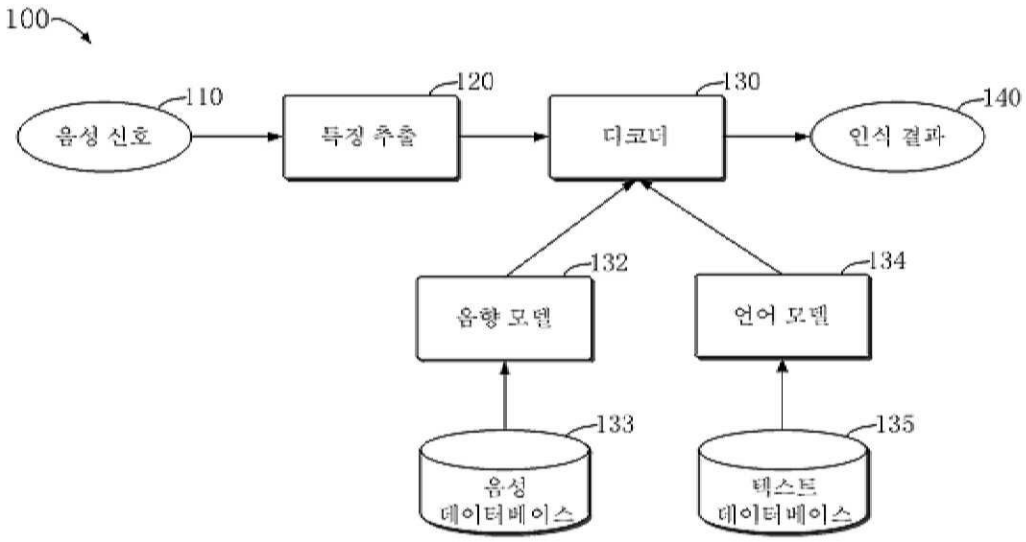
[0068] 본 개시의 전반 서술에서, 기계 판독 가능 매체는 유형의 매체일 수 있는바, 이는, 명령어 실행 시스템, 장치 또는 기기에 사용하기 위한, 또는 명령어 실행 시스템, 장치 또는 기기와 결합하여 사용하기 위한 프로그램을 포함 또는 저장할 수 있다. 기계 판독 가능 매체는 기계 판독 가능 신호 매체 또는 기계 판독 가능 저장 매체일 수 있다. 기계 판독 가능 매체는 전자의, 자성의, 광학의, 전자기의, 적외선의 또는 반도체의 시스템, 장치 또는 기기, 또는 상술한 내용의 임의의 적합한 조합을 포함할 수 있으나 이에 한정되지 않는다. 기계 판독 가능 저장 매체의 더 구체적인 예시는 하나 또는 복수의 선에 기반하는 전기 연결, 휴대형 컴퓨터 디스크, 하드 디스크, 랜덤 액세스 메모리(RAM), 읽기 전용 메모리(ROM), 소거 및 프로그램 가능 읽기 전용 메모리(EPROM 또는 플래시 메모리), 광섬유, 휴대용 콤팩트 디스크 읽기 전용 메모리(CD-ROM), 광학 저장 기기, 자기 저장 기기, 또는 상술한 내용의 임의의 적합한 조합을 포함하게 된다.

[0069] 이 외에, 비록 각 동작 또는 단계를 묘사함에 있어서 특정 순서를 적용하였지만, 이러한 동작 또는 단계가 도시된 특정 순서로 또는 순차적 순서로 실행되어야 하거나, 또는 모든 도시한 동작 또는 단계가 실행되어야 하는 것으로, 이로써 기대하는 결과를 취득하기 위한 것임을 이해하여야 한다. 일정한 환경에서는, 다중 태스크 및 병렬 처리가 유익할 수 있다. 마찬가지로, 비록 상기의 서술에는 약간의 구체 구현 디테일이 포함되지만, 이러한 것은 본 개시의 범위에 대한 제한으로 해석되어서는 안된다. 단독의 실시예의 전반 서술에서 설명되는 어떤 특징 또한, 조합 가능하게 하나의 구현에 구현될 수 있다. 반대로, 하나의 구현의 전반 서술에서 설명되는 여러 가지 특징도 단독으로 또는 임의의 적합한 서브조합의 방식으로 복수의 실시예에 구현될 수 있다.

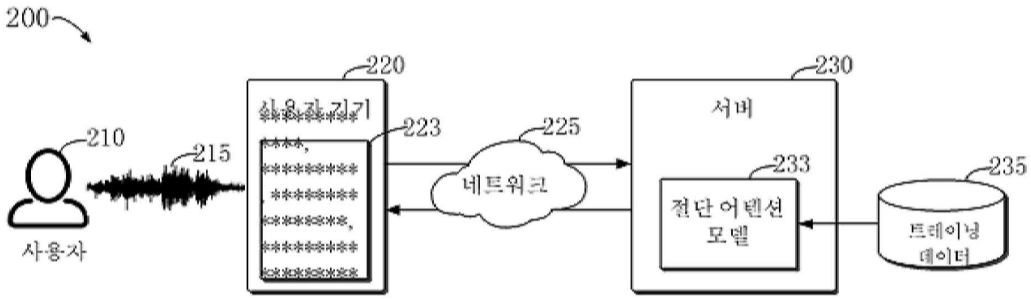
[0070] 비록 이미 구조적 특징 및/또는 방법론적 동작에 특정된 언어를 적용하여 본 개시의 실시예에 대해 설명하였지만, 첨부된 특허청구범위가 한정하는 주제는 위에서 설명한 특정 특징 또는 동작에 한정되는 것이 아님을 이해하여야 한다. 반대로, 위에서 설명한 특정 특징 및 동작은 단지 특허청구범위의 예시 형식을 구현하는 것일 뿐이다.

도면

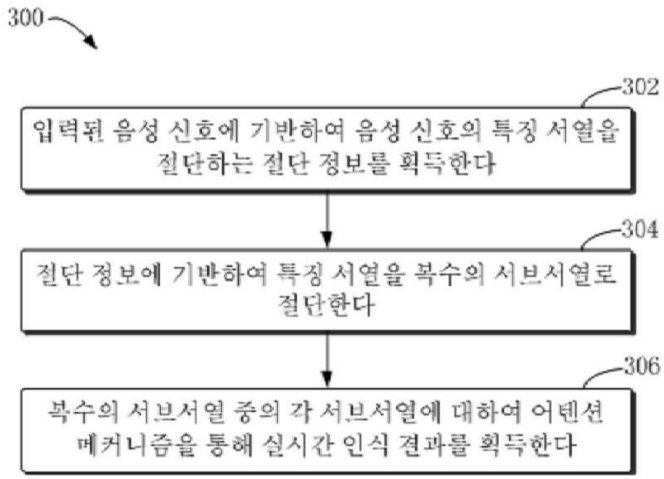
도면1



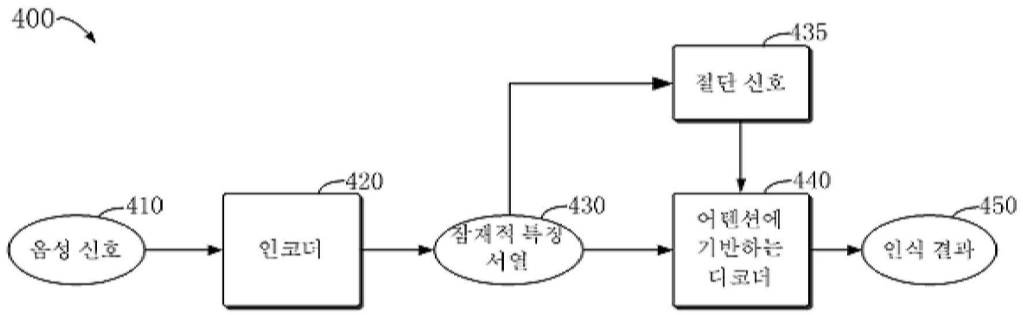
도면2



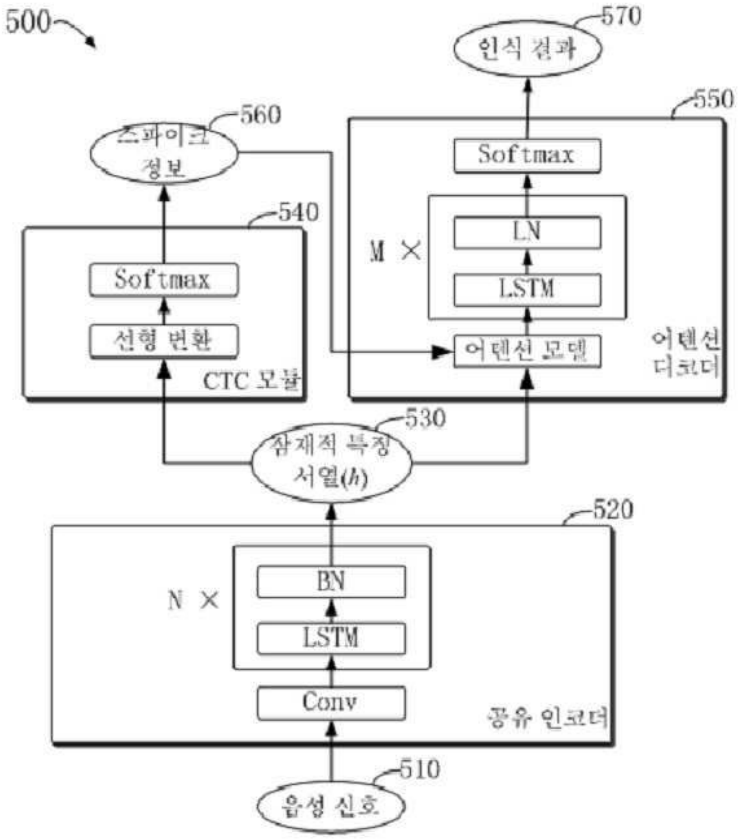
도면3



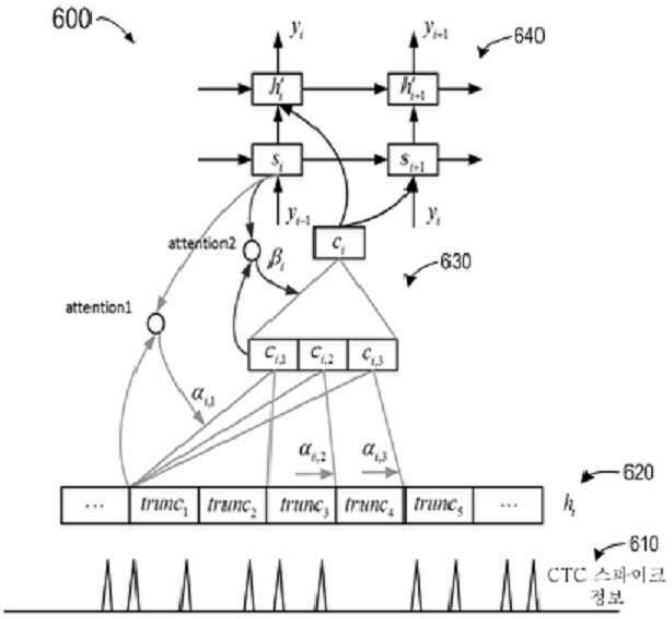
도면4


400
410
음성 신호
420
인코더
430
잠재적 특징 서열
435
절단 신호
440
어텐션에 기반하는 디코더
450
인식 결과

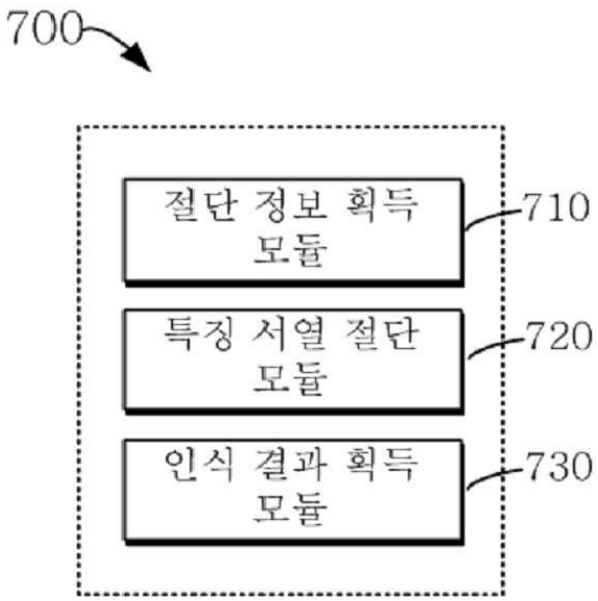
도면5


500
570
인식 결과
550
560
스파이크 정보
Softmax
M ×
LN
LSTM
어텐션 모델
어텐션 디코더
540
Softmax
선형 변환
CTC 모듈
530
잠재적 특징 서열(h)
520
N ×
BN
LSTM
Conv
공유 인코더
510
음성 신호

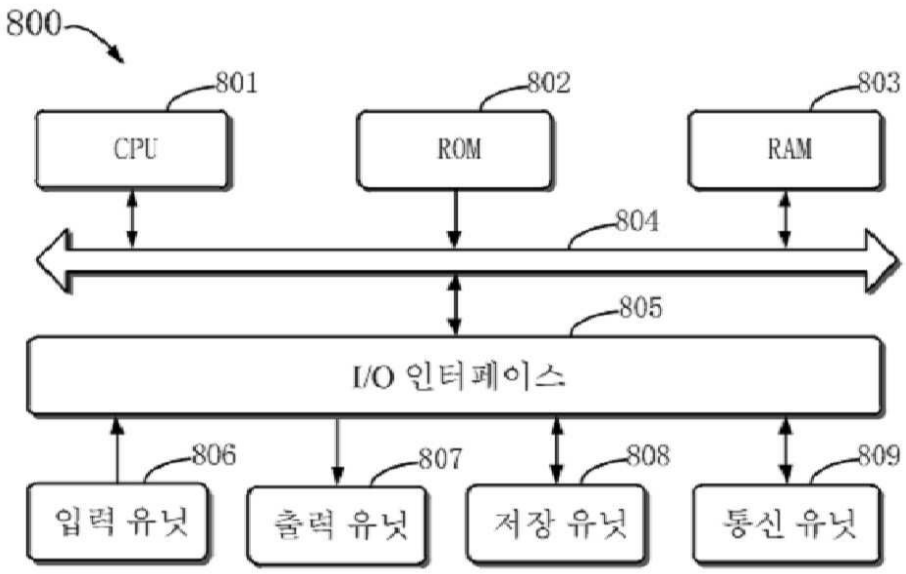
도면6


600
640
attention2
attention1
630
620
610
CTC 스파이크 정보

도면7


700
절단 정보 획득 모듈
710
특징 서열 절단 모듈
720
인식 결과 획득 모듈
730

도면8


800
801
CPU
802
ROM
803
RAM
804
805
I/O 인터페이스
806
입력 유닛
807
출력 유닛
808
저장 유닛
809
통신 유닛