(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2018-0098806
(43) 공개일자 2018년09월05일

(51) 국제특허분류(Int. Cl.)
G10L 25/93 (2013.01) *G10L 15/06* (2006.01)
G10L 19/02 (2006.01) *G10L 19/04* (2006.01)
G10L 25/30 (2013.01) *G10L 25/78* (2013.01)
(52) CPC특허분류
G10L 25/93 (2013.01)
G10L 15/06 (2013.01)
(21) 출원번호 10-2017-0025397
(22) 출원일자 2017년02월27일
심사청구일자 2018년02월02일

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
(72) 발명자
김현우
대전광역시 서구 탄방로 10 (탄방동)
정호영
대전광역시 서구 둔산북로 160, 8동 1302호 (둔산동, 한마루아파트)
(뒷면에 계속)
(74) 대리인
특허법인지명

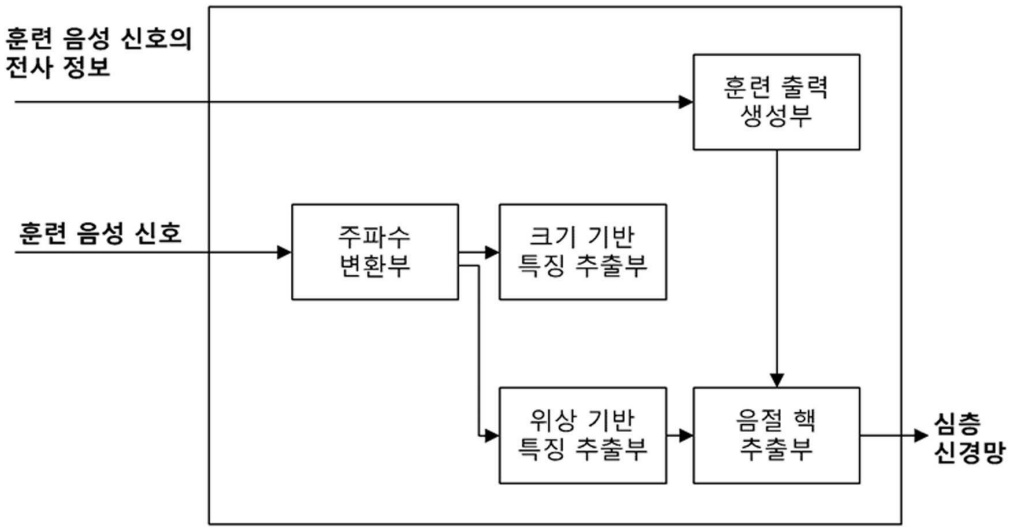
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **자연어 인식 성능 개선 방법 및 장치**

(57) 요 약

본 발명의 일면에 따른 자연어 인식 성능 개선 장치는 음성 신호를 복수의 프레임으로 분할하고, 이산 푸리에 변환을 적용하여 시간 영역에서 주파수 영역으로 변환하는 주파수 변환부; 상기 주파수 영역으로 변환된 음성 신호의 크기로부터 크기 특징을 추출하는 크기 특징 추출부; 상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 위상 특징을 추출하는 위상 특징 추출부; 상기 크기 특징과 상기 위상 특징을 심층 신경망의 입력으로 하고, 음절 핵을 검출하는 음절 핵 검출부; 상기 입력 음성 신호로부터 음성 구간과 비음성 구간을 검출하는 음성 검출부; 상기 검출된 음절 핵과 상기 검출된 음성 구간의 시간을 이용하여 발성 속도를 결정하는 발성 속도 결정부; 상기 발성 속도를 이용하여 시간축 변환 정도를 계산하는 계산부; 및 상기 시간축 변환 정도를 이용하여 음향 모델에 적합한 음성의 길이로 변환하는 시간축 변환부;를 포함한다.

대 표 도 - 도3



(52) CPC특허분류
G10L 19/02 (2013.01)
G10L 19/04 (2013.01)
G10L 25/30 (2013.01)
G10L 25/78 (2013.01)
(72) 발명자
박전규
대전광역시 유성구 대덕대로541번길 68, 103동 205호 (도룡동, 현대아파트)
이윤근
대전광역시 서구 청사서로 11, 103동 1406호 (월평동, 무지개아파트)

이 발명을 지원한 국가연구개발사업
과제고유번호 R0126-15-1117
부처명 미래창조과학부
연구관리전문기관 정보통신기술진흥센터
연구사업명 SW컴퓨팅산업원천기술개발
연구과제명 언어학습을 위한 자유발화형 음성대화처리 원천기술 개발
기 여 율 1/1
주관기관 한국전자통신연구원
연구기간 2016.03.01 ~ 2017.02.28

명 세 서

청구범위

청구항 1

음성 신호를 복수의 프레임으로 분할하고, 이산 푸리에 변환을 적용하여 시간 영역에서 주파수 영역으로 변환하는 주파수 변환부;

상기 주파수 영역으로 변환된 음성 신호의 크기로부터 크기 특징을 추출하는 크기 특징 추출부;

상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 위상 특징을 추출하는 위상 특징 추출부;

상기 크기 특징과 상기 위상 특징을 심층 신경망의 입력으로 하고, 음절 핵을 검출하는 음절 핵 검출부;

상기 입력 음성 신호로부터 음성 구간과 비음성 구간을 검출하는 음성 검출부;

상기 검출된 음절 핵과 상기 검출된 음성 구간의 시간을 이용하여 발성 속도를 결정하는 발성 속도 결정부;

상기 발성 속도를 이용하여 시간축 변환 정도를 계산하는 계산부; 및

상기 시간축 변환 정도를 이용하여 음향 모델에 적합한 음성의 길이로 변환하는 시간축 변환부;

를 포함하는 자연어 인식 성능 개선 장치.

청구항 2

제1항에 있어서,

상기 크기 특징은 멜-스케일 필터뱅크 로그 에너지, MFCC, LPC, 피치, 하모닉 성분, 스펙트럼 평탄도 중 적어도 하나를 포함하는 것

인 자연어 인식 성능 개선 장치.

청구항 3

제1항에 있어서,

상기 위상 특징은 델타-위상 스펙트럼, 위상 왜곡 편차, 그룹 지연, 순환 분산 중 적어도 하나를 포함하는 것

인 자연어 인식 성능 개선 장치.

청구항 4

제1항에 있어서,

상기 시간축 변환 정도는 변화율 또는 중첩률 중 어느 하나인 것

인 자연어 인식 성능 개선 장치.

청구항 5

제1항에 있어서,

상기 음성 검출부는,

깨끗한 음성과 잡음의 DFT 계수 분포를 정규 분포로 모델링하고 우도비 검정을 수행하는 것

인 자연어 인식 성능 개선 장치.

청구항 6

제1항에 있어서,

상기 음절 핵 검출부에서 사용하는 심층 신경망은,

훈련 음성 신호와, 훈련 음성 신호의 전사 정보를 입력으로 하고,

상기 훈련 음성 신호를 주파수 영역으로 변환하여 크기 특징 및 위상 특징을 추출하고,

상기 훈련 음성 신호의 전사 정보로부터 음소의 분류 항목을 다중 프레임 출력으로 구성하고,

상기 크기 특징, 위상 특징을 입력으로 하고 상기 다중 프레임 출력으로 구성된 음소의 분류 항목을 출력으로 하는 심층 신경망을 훈련하고, 크로스 엔트로피를 기준치로 하여 역전파 알고리즘으로 훈련하는 것

인 자연어 인식 성능 개선 장치.

청구항 7

제6항에 있어서,

상기 음소의 분류 항목은

묵음, 자음, 음절 핵 및 연속 음절 핵을 포함하는 것

인 자연어 인식 성능 개선 장치.

청구항 8

제6항에 있어서,

상기 다중 프레임 출력은

음성 신호의 전사 정보와 음성 인식기를 사용하여 강제 정렬을 수행함으로써 음소의 분류 항목에 해당하는 음성 신호 구간을 추정하고, 이웃 프레임들의 음소의 분류 항목을 묶어 다중 프레임 출력하는 것

인 자연어 인식 성능 개선 장치.

청구항 9

(1) 소정의 시간 간격으로 음성 신호를 복수의 프레임으로 분할하고 이산 푸리에 변환을 적용하여 시간 영역에서 주파수 영역으로 변환하는 단계;

(2) 상기 주파수 영역으로 변환된 음성 신호의 크기로부터 크기 특징을 추출하는 단계;

(3) 상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 위상 특징을 추출하는 단계;

(4) 상기 크기 특징과 위상 특징을 심층 신경망의 입력으로 사용하여 음절 핵을 검출하는 단계;

(5) 상기 입력 음성 신호로부터 음성 구간과 비음성 구간을 검출하는 단계;

(6) 상기 검출된 음절 구간의 시간을 이용하여 발성 속도를 결정하는 단계;

(7) 상기 발성 속도를 이용하여 시간축 변환 정도를 계산하는 단계; 및

(8) 상기 시간축 변환 정도를 이용하여 음향 모델에 적합하게 음성의 길이를 변환하는 단계;

를 포함하는 자연어 인식 개선 방법.

청구항 10

제9항에 있어서,

상기 크기 특징을 추출하는 단계는,

상기 크기 특징으로서, 멜-스케일 필터뱅크 로그 에너지, MFCC, LPC, 피치, 하모닉 성분, 스펙트럼 평탄도를 추출하는 단계인 것

인 자연어 인식 성능 개선 방법.

청구항 11

제9항에 있어서,

상기 위상 특징을 추출하는 단계는,

상기 위상 특징으로서, 델타-위상 스펙트럼, 위상 왜곡 편차, 그룹 지연, 순환 분산를 추출하는 단계인 것

인 자연어 인식 성능 개선 방법.

청구항 12

제9항에 있어서,

시간축 변환 정도를 계산하는 단계는,

상기 시간축 변환 정도로서, 변화율 또는 중첩률 중 어느 하나를 계산하는 단계인 것

인 자연어 인식 성능 개선 방법.

청구항 13

제9항에 있어서,

음성 구간과 비음성 구간을 검출하는 단계는,

깨끗한 음성과 잡음의 DFT 계수 분포를 정규 분포로 모델링하고 우도비 검정을 수행하는 단계인 것

인 자연어 인식 성능 개선 방법.

청구항 14

제9항에 있어서,

상기 음절 핵을 검출하는 단계는,

상기 심층 신경망의 입력으로서, 훈련 음성 신호와, 훈련 음성 신호의 전사 정보를 이용하고,

상기 훈련 음성 신호를 주파수 영역으로 변환하여 크기 특징 및 위상 특징을 추출하고,

상기 훈련 음성 신호의 전사 정보로부터 음소의 분류 항목을 다중 프레임 출력으로 구성하고,

상기 크기 특징, 위상 특징을 입력으로 하고 상기 다중 프레임 출력으로 구성된 음소의 분류 항목을 출력으로 하는 심층 신경망을 훈련하고, 크로스 엔트로피를 기준치로 하여 역전파 알고리즘으로 훈련하는 단계인 것

인 자연어 인식 성능 개선 방법.

청구항 15

제14항에 있어서,

상기 음소의 분류 항목은,

묵음, 자음, 음절 핵 및 연속 음절 핵을 포함하는 것

인 자연어 인식 성능 개선 방법.

청구항 16

제14항에 있어서,

상기 다중 프레임 출력은

음성 신호의 전사 정보와 음성 인식기를 사용하여 강제 정렬을 수행함으로써 음소의 분류 항목에 해당하는 음성 신호 구간을 추정하고, 이웃 프레임들의 음소의 분류 항목을 묶어 다중 프레임 출력하는 것

인 자연어 인식 성능 개선 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 자연어 인식 성능 개선 방법 및 장치에 관한 것으로 구체적으로, 다양한 발성 속도를 갖는 자연어 인식 성능을 향상시키는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 일반적으로 자연어(spontaneous speech)에서는 다양한 발성 속도(speaking-rate)가 관찰되는데, 적절 속도로 발성한 음성으로 학습된 음성 인식기에서는 자연어 인식 성능이 떨어지게 된다. 이러한 발성 속도의 변이를 대처하는 방법으로 특징이나 신호 영역에서 음향 모델에 적합한 음성의 길이를 조절하는 방법이 있다.

[0003] 예컨대, 특징 영역에서 켑스트럼 길이 정규화(cepstrum length normalization) 방법과 신호 영역에서 PSOLA(Pitch Synchronous Overlap and Add) 기반 시간축 변환(time scale modification) 방법이 있다. 켑스트럼 길이 변화율이나 PSOLA의 중첩률 조절을 위해서 발성 속도의 측정이 선행되어야 한다.

[0004] 일정 시간 동안 발성한 음절 수를 추정하는 방법으로 발성 속도를 결정할 수 있다. 음절은 대부분 모음으로 구성된 음절 핵(syllabic nucleus)을 포함하고 있는데, 음절 핵은 두음(onset)과 말음(code)에 비해 에너지와 주기성이 큰 성질을 가지고 있어서, 음절 책과 다른 음절 핵 사이에서 에너지와 주기성이 약해지거나 사라지고 음절 핵에서 강해지는 현상 발생한다. 음절 핵에서 에너지와 주기성의 정점이 형성되기 때문에, 에너지와 주기성을 이용하여 음절 핵을 검출하고, 정점의 개수를 음절 수로 사용한다.

[0005] 구체적으로 음성 신호를 복수의 프레임으로 분할하고, 매 프레임마다 에너지 관련 특징(전대역 에너지, 부대역 에너지, 포락선의 상관도, 저대역 변조 에너지 등)과 주기성 관련 특징(피치, 하모닉 성분 크기)을 추출한 후, 특징의 정점을 검출하여 정점의 개수를 음성 구간의 길이로 나누어 발성 속도를 결정한다. 그러나, 선행 기술에 따르면, '과일', '거의', '수입'과 같이 음절 핵과 음절 핵이 직접 이어지거나, 음절 핵 사이에 두음과 말음으로 공명음('ㄴ', 'ㅁ', 'ㄹ', 'ㅇ')이 존재할 때, 음절 핵과 다른 음절 핵 사이에서 에너지와 주기성이 약해지거나 사라졌다가 다시 강해지는 현상이 발생하지 않기 때문에 에너지와 주기성의 정점 검출이 어려운 문제가 있다.

[0006] 최근 활발하게 연구되는 심층 신경망(Deep Neural Network)은 입력 계층과 출력 계층 사이에 다수의 은닉 계층들로 이루어진 신경망으로 입력과 출력의 복잡한 관계를 표현한다. 특히 입력 신호의 프레임간 동적 정보를 활

용하고, 암시적인 입력 신호의 특징을 추출함으로써 출력과의 관계를 정교하게 표현해주는 장점이 있다. 이러한 장점은 음절 핵이 이어지거나 음절 핵 사이에서 공명음이 존재할 때, 음절 핵을 검출하기 어려운 문제를 해결할 수 있다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 문제를 해결하기 위하여, 자연어 인식 성능 개선 방법 및 장치를 제공하는 것을 그 목적으로 한다.

[0008] 본 발명에서는 심층 신경망 기반으로 음절 핵을 검출하고, 발성 속도에 따라 길이 변화율 또는 중첩율을 조절하여 자연어 인식 성능을 개선하는 방법을 제공한다. 심층 신경망의 성능을 높이기 위하여, 크기 특징 외에 위상 특징을 입력으로 사용하고 다중 프레임 출력을 사용하는 방법을 제공한다.

[0009] 본 발명의 목적은 다양한 발성 속도를 갖는 자연어 인식 성능을 향상시키기 위하여, 심층 신경망을 토대로 발성 속도를 결정하고, 길이 변화율 또는 중첩율을 조절하는 방법 및 장치를 제공하는 것이다. 주파수 영역으로 변환된 음성 신호의 크기 특징 외에 위상 특징을 추출하고, 다중 프레임 출력을 사용하는 심층 신경망을 토대로 음절 핵을 검출하고, 음성 검출기로 검출된 음성 구간의 시간으로 음절 핵의 개수를 발성 속도를 결정하고, 발성 속도에 따라 길이 변화율 또는 중첩률을 계산하고, 음향 모델에 적합한 음성의 길이로 켑스트럼 길이 정규화 또는 시간축 변환을 수행함으로써 자연어 인식 성능을 개선하는 방법 및 장치를 제공하는 것을 목적으로 한다.

[0010] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0011] 전술한 목적을 달성하기 위한 본 발명의 일 측면에 따른 자연어 인식 성능 개선 장치는, 음성 신호를 복수의 프레임으로 분할하고, 이산 푸리에 변환을 적용하여 시간 영역에서 주파수 영역으로 변환하는 주파수 변환부; 상기 주파수 영역으로 변환된 음성 신호의 크기로부터 크기 특징을 추출하는 크기 특징 추출부; 상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 위상 특징을 추출하는 위상 특징 추출부; 상기 크기 특징과 상기 위상 특징을 심층 신경망의 입력으로 하고, 음절 핵을 검출하는 음절 핵 검출부; 상기 입력 음성 신호로부터 음성 구간과 비음성 구간을 검출하는 음성 검출부; 상기 검출된 음절 핵과 상기 검출된 음성 구간의 시간을 이용하여 발성 속도를 결정하는 발성 속도 결정부; 상기 발성 속도를 이용하여 시간축 변환 정도를 계산하는 계산부; 및 상기 시간축 변환 정도를 이용하여 음향 모델에 적합한 음성의 길이로 변환하는 시간축 변환부를 포함할 수 있다.

[0012] 상기 크기 특징은 멜-스케일 필터뱅크 로그 에너지, MFCC, LPC, 피치, 하모닉 성분, 스펙트럼 평탄도 중 적어도 하나를 포함한다.

[0013] 상기 위상 특징은 델타-위상 스펙트럼, 위상 왜곡 편차, 그룹 지연, 순환 분산 중 적어도 하나를 포함한다.

[0014] 상기 시간축 변환 정도는 변화율 또는 중첩률 중 어느 하나일 수 있다.

[0015] 상기 음성 검출부는, 깨끗한 음성과 잡음의 DFT 계수 분포를 정규 분포로 모델링하고 우도비 검정을 수행한다.

[0016] 상기 음절 핵 검출부에서 사용하는 심층 신경망은, 훈련 음성 신호와, 훈련 음성 신호의 전사 정보를 입력으로 하고, 상기 훈련 음성 신호를 주파수 영역으로 변환하여 크기 특징 및 위상 특징을 추출하고, 상기 훈련 음성 신호의 전사 정보로부터 음소의 분류 항목을 다중 프레임 출력으로 구성하고, 상기 크기 특징, 위상 특징을 입력으로 하고 상기 다중 프레임 출력으로 구성된 음소의 분류 항목을 출력으로 하는 심층 신경망을 훈련하고, 크로스 엔트로피를 기준치로 하여 역전파 알고리즘으로 훈련한다.

[0017] 상기 음소의 분류 항목은, 묵음, 자음, 음절 핵 및 연속 음절 핵을 포함한다.

[0018] 상기 다중 프레임 출력은, 음성 신호의 전사 정보와 음성 인식기를 사용하여 강제 정렬을 수행함으로써 음소의 분류 항목에 해당하는 음성 신호 구간을 추정하고, 이웃 프레임들의 음소의 분류 항목을 묶어 다중 프레임 출력

하는 것이다.

[0019] 한편, 본 발명의 다른 측면에 따른 자연어 인식 성능 개선 방법은, (1) 소정의 시간 간격으로 음성 신호를 복수의 프레임으로 분할하고 이산 푸리에 변환을 적용하여 시간 영역에서 주파수 영역으로 변환하는 단계; (2) 상기 주파수 영역으로 변환된 음성 신호의 크기로부터 크기 특징을 추출하는 단계; (3) 상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 위상 특징을 추출하는 단계; (4) 상기 크기 특징과 위상 특징을 심층 신경망의 입력으로 사용하여 음절 핵을 검출하는 단계; (5) 상기 입력 음성 신호로부터 음성 구간과 비음성 구간을 검출하는 단계; (6) 상기 검출된 음절 구간의 시간을 이용하여 발성 속도를 결정하는 단계; (7) 상기 발성 속도를 이용하여 시간축 변환 정도를 계산하는 단계; 및 (8) 상기 시간축 변환 정도를 이용하여 음향 모델에 적합하게 음성의 길이를 변환하는 단계를 포함할 수 있다.

[0020] 상기 크기 특징을 추출하는 단계는, 상기 크기 특징으로서, 멜-스케일 필터뱅크 로그 에너지, MFCC, LPC, 피치, 하모닉 성분, 스펙트럼 평탄도를 추출하는 단계를 포함한다.

[0021] 상기 위상 특징을 추출하는 단계는, 상기 위상 특징으로서, 델타-위상 스펙트럼, 위상 왜곡 편차, 그룹 지연, 순환 분산를 추출하는 단계이다.

[0022] 시간축 변환 정도를 계산하는 단계는, 상기 시간축 변환 정도로서, 변화율 또는 중첩률 중 어느 하나를 계산하는 단계이다.

[0023] 음성 구간과 비음성 구간을 검출하는 단계는, 깨끗한 음성과 잡음의 DFT 계수 분포를 정규 분포로 모델링하고 우도비 검정을 수행하는 단계를 포함한다.

[0024] 상기 음절 핵을 검출하는 단계는, 상기 심층 신경망의 입력으로서, 훈련 음성 신호와, 훈련 음성 신호의 전사 정보를 이용하고, 상기 훈련 음성 신호를 주파수 영역으로 변환하여 크기 특징 및 위상 특징을 추출하고, 상기 훈련 음성 신호의 전사 정보로부터 음소의 분류 항목을 다중 프레임 출력으로 구성하고, 상기 크기 특징, 위상 특징을 입력으로 하고 상기 다중 프레임 출력으로 구성된 음소의 분류 항목을 출력으로 하는 심층 신경망을 훈련하고, 크로스 엔트로피를 기준치로 하여 역전파 알고리즘으로 훈련하는 단계이다.

[0025] 상기 음소의 분류 항목은, 묵음, 자음, 음절 핵 및 연속 음절 핵을 포함한다.

[0026] 상기 다중 프레임 출력은, 음성 신호의 전사 정보와 음성 인식기를 사용하여 강제 정렬을 수행함으로써 음소의 분류 항목에 해당하는 음성 신호 구간을 추정하고, 이웃 프레임들의 음소의 분류 항목을 묶어 다중 프레임 출력하는 단계이다.

발명의 효과

[0027] 본 발명에 따르면, 심층 신경망을 이용하여 발성 속도를 결정하고 길이 변화율 또는 중첩률을 조절하여 다양한 발성 속도를 갖는 자연어에 대한 인식 성능을 개선할 수 있다. 심층 신경망 기반의 음절 핵 검출에서 크기 특징 외에 위상 특징을 입력으로 사용하고 다중 프레임 출력을 사용함으로써, 음절 핵이 이어지거나 음절 핵 사이에서 공명음이 존재하더라도 효과적으로 음절 핵을 검출할 수 있다. 본 발명에 따르면, 발성 속도 결정의 정확도를 높일 수 있고, 음절 핵 사이의 시간을 측정함으로써 장음화 현상을 검출하는데에도 효과가 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명에 따른 자연어 인식 성능 개선 방법이 구현되는 컴퓨터 시스템의 구성을 설명하기 위한 예시도.

도 2는 본 발명에 따른 자연어 인식 성능 개선 장치의 일실시예를 나타내는 구성도.

도 3은 훈련 음성 신호로부터 심층 신경망을 훈련하는 장치의 일실시예를 나타내는 구성도.

도 4는 인공 신경망의 예측 방법을 설명하기 위한 예시도.

도 5는 본 발명의 일실시예에 따른 자연어 인식 성능 개선 방법을 나타내는 흐름도.

도 6는 본 발명의 일실시예에 따른 훈련 음성 신호로부터 심층 신경망을 훈련하는 방법을 나타내는 흐름도.

발명을 실시하기 위한 구체적인 내용

[0029] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 것이며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하며, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다. 한편, 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성소자, 단계, 동작 및/또는 소자는 하나 이상의 다른 구성소자, 단계, 동작 및/또는 소자의 존재 또는 추가를 배제하지 않는다.

[0030] 이하, 본 발명의 바람직한 실시예에 대하여 첨부한 도면을 참조하여 상세히 설명하기로 한다.

[0031] 도 1은 본 발명에 따른 자연어 인식 성능 개선 방법이 구현되는 컴퓨터 시스템의 구성을 설명하기 위한 예시도이다.

[0032] 한편, 본 발명의 실시예에 따른 자연어 인식 성능 개선 방법은 컴퓨터 시스템에서 구현되거나, 또는 기록매체에 기록될 수 있다. 도 1에 도시된 바와 같이, 컴퓨터 시스템은 적어도 하나 이상의 프로세서(110)와, 메모리(120)와, 사용자 입력 장치(150)와, 데이터 통신 버스(130)와, 사용자 출력 장치(160)와, 저장소(140)를 포함할 수 있다. 전술한 각각의 구성 요소는 데이터 통신 버스(130)를 통해 데이터 통신을 한다.

[0033] 컴퓨터 시스템은 네트워크(180)에 연결된 네트워크 인터페이스(170)를 더 포함할 수 있다. 상기 프로세서(110)는 중앙처리 장치(central processing unit (CPU))이거나, 혹은 메모리(120) 및/또는 저장소(140)에 저장된 명령어를 처리하는 반도체 장치일 수 있다.

[0034] 상기 메모리(120) 및 상기 저장소(140)는 다양한 형태의 휘발성 혹은 비휘발성 저장매체를 포함할 수 있다. 예컨대, 상기 메모리(120)는 ROM(123) 및 RAM(126)을 포함할 수 있다.

[0035] 따라서, 본 발명의 실시예에 따른 자연어 인식 성능 개선 방법은 컴퓨터에서 실행 가능한 방법으로 구현될 수 있다. 본 발명의 실시예에 따른 자연어 인식 성능 개선 방법이 컴퓨터 장치에서 수행될 때, 컴퓨터로 판독 가능한 명령어들이 본 발명에 따른 운영 방법을 수행할 수 있다.

[0036] 한편, 상술한 본 발명에 따른 자연어 인식 성능 개선 방법은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현되는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록 매체로는 컴퓨터 시스템에 의하여 해독될 수 있는 데이터가 저장된 모든 종류의 기록 매체를 포함한다. 예를 들어, ROM(Read Only Memory), RAM(Random Access Memory), 자기 테이프, 자기 디스크, 플래시 메모리, 광 데이터 저장장치 등이 있을 수 있다. 또한, 컴퓨터로 판독 가능한 기록매체는 컴퓨터 통신망으로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 읽을 수 있는 코드로서 저장되고 실행될 수 있다.

[0037] 도 2는 본 발명에 따른 자연어 인식 성능 개선 장치의 일실시예를 나타내는 구성도이다.

[0038] 도 2에 따르면, 본 발명에 따른 자연어 인식 성능 개선 장치는 종래의 기술과 달리, 음성 신호의 동적 정보를 활용하여 음절 핵을 검출하기 위하여 심층 신경망을 토대로 발성 속도를 결정한다.

[0039] 본 발명에 따른 자연어 인식 성능 개선 장치는 주파수 변환부, 특징 추출부, 음절 핵 검출부, 음성 검출부, 발성 속도 결정부, 길이 변화율 및 중첩률 계산부, 길이 정규화 및 시간축 변환부를 포함한다.

[0040] 주파수 변환부는 소정의 시간 간격(예컨대, 30ms)으로 음성신호를 복수의 프레임으로 분할하고, 이산 푸리에 변환(Discrete Fourier Transform, DFT) 을 적용하여 시간 영역에서 주파수 영역으로 변환한다. 여기서, 통상적인 푸리에 변환은 다음 식과 같다.

수학식 1

$$F(w) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-iwt}dt$$

[0041]

[0042] 이산 푸리에 변환은 다음 식과 같다.

수학식 2

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn}$$

[0043] 단, $k = 0, 1, \ldots, N-1$

[0044] 이산 푸리에 변환은 이산적인 신호에 대하여, 푸리에 변환과 같은 효과를 얻을 수 있다. 이산 푸리에 변환을 처리하기 위하여, 쿨리-튜키 알고리즘 또는 프라임 팩터 알고리즘, 브룬 알고리즘, 레이더 알고리즘, 블루스타인 알고리즘 등의 고속 푸리에 변환 알고리즘을 이용할 수 있다.

[0045] 특징 추출부는 크기 특징 추출부; 및 위상 특징 추출부;를 포함한다.

[0046] 크기 특징 추출부는 상기 주파수 영역으로 변환된 음성 신호의 크기로부터 멜-스케일 필터뱅크 로그 에너지(mel-scale filterbank log energy), MFCC(Mel-Frequency Cepstral Coefficient), LPC(Linear Prediction Coefficient), 피치, 하모닉 성분, 스펙트럼 평탄도(spectral flatness) 등의 정보를 추출한다.

[0047] 위상 특징 추출부는 상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 특징을 추출한다. 음성 신호의 위상은 직접적이고 접근 가능한 정보를 명시적으로 보여주지 못하기 때문에 일반적으로 음성 신호 처리 분야에서 사용되지 아니하나, 심층 신경망을 통하여 분석하는 경우, 입력 신호의 암시적인 정보를 추출할 수 있는 장점이 있어, 본 발명에서는 위상 특징을 크기 특징과 함께 사용한다.

[0048] 상기 추출된 위상 특징은 델타-위상 스펙트럼(delta-phase spectrum), 위상 왜곡 편차(phase distortion deviation), 그룹 지연(group delay), 순환 분산(circular variance)를 포함한다.

[0049] 음절 핵 검출부는 상기 크기 특징과 위상 특징을 심층 신경망의 입력으로 사용하여 음절 핵을 검출한다. 여기서 심층 신경망은 사전에 훈련 음성 신호로부터 획득한다.

[0050] 음성 검출부는 입력 음성 신호로부터 음성 구간과 비음성 구간을 검출한다. 예컨대, 깨끗한 음성과 잡음의 DFT 계수 분포를 정규 분포로 모델링하여 우도비 검정(Likelihood Ratio Test, LRT)을 수행하여 음성 구간을 검출한다.

[0051] 발성 속도 결정부는 상기 검출된 음절 핵의 음절 수(fram_count)로 사용하고, 상기 음성 검출기로 검출된 음성 구간의 시간 구간(speech_interval)으로 음절 수를 나누어 발성 속도(rate)를 결정한다. 발성 속도는 음절 시간 간격이 일정하다는 전제에서 일정 길이 이상의 음성 구간이 존재해야 측정할 수 있다.

수학식 3

[0052] $$\text{rate} = \frac{fram\_count}{speech\_interval}$$

[0053] 길이 변화율 및 중첩률 계산부는 상기 발성 속도를 사용하여 시간축 변환의 정도를 나타내는 길이 변화율 및 중첩률을 계산한다. 예컨대, PSOLA의 중첩률(factor)을 아래와 같이 선형적으로 조절할 수 있다.

수학식 4

[0054] $$\text{factor} = \text{MIN}(\text{MAX}(\beta * (\text{rate} - \gamma) + 1.0, 1.0), 1.5)$$

[0055] 이때, 느린 속도로 발성하는 자연어의 인식기 성능은 크게 떨어지지 않기 때문에 시간축 변환을 적용하지 않는다. 또한 일정 범위를 벗어난 중첩률을 사용하여 PSOLA 기법을 적용하면 부자연스러운 합성음이 되어 인식 성능이 저하되기 때문에 최대 중첩률을 1.5로 제한한다.

[0056] 길이 정규화 및 시간축 변환부는 상기 길이 변화율 또는 중첩률을 사용하여 음향 모델에 적합한 음성의 길이로 켑스트럼 길이 정규화 또는 시간축 변환을 수행한다. 예컨대, 기존의 PSOLA 기반 시간축 변환을 사용한다. PSOLA 기법은 피치 단위로 분석된 음성을 중첩해서 더하는 알고리즘으로 합성된 신호의 정점 간격이 일치하기 때문에 피치의 변화가 발생하지 않는다.

[0057] 도 3은 훈련 음성 신호로부터 심층 신경망을 훈련하는 장치의 일 실시예를 나타내는 구성도이다.

[0058] 주파수 변환부는 훈련 음성 신호를 주파수 영역으로 변환한다.

[0059] 특징 추출부는 상기 변환된 훈련 음성 신호로부터 크기 특징과 위상 특징을 추출한다. 훈련 출력 생성부는 훈련 음성 신호의 전사 정보로부터 심층 신경망에 사용할 출력을 생성한다. 훈련 음성 신호의 전사 정보에서 획득한 음소를 "묵음", "자음", "음절 핵", "연속 음절 핵" 4가지로 분류하고, 각 분류 항목에 "1", "2", "3", "4"의 번호를 부여하여 심층 신경망 모델의 출력으로 사용한다. 이는 예시적인 것이며, 분류의 종류, 분류 방법을 한정하는 것은 아니다.

[0060] 예컨대, 음성 인식 분야에서 널리 사용되는 TIMIT 훈련 음성 데이터 베이스는 발음 기준으로 61개의 음소가 존재하고, 이에 대하여 "묵음", "자음", "음절 핵"을 분류한 것을 아래의 표에 나타내었다.

표 1

[0061]

출력	분류	음소
1	묵음	h#, epi, pau
2	자음	jh, ch, b, d, g, p, t, k, dx, s, sh, z, zh, f, th, v, dh, m, n, nx, ng, l, r, w, y, hh, hv, q, bcl, dcl, gcl, pcl ,tcl, kcl
3	음절 핵	ae, aa, ah, eh, iy, ih, uh, uw, aw, ay, ey, oy, ow, ao, ax, ax-h, ix, ux, er, axr, el, em, en, eng

[0062] "연속 음절 핵"은 음절 핵이 연이어 나올 때, 뒤의 음절 핵을 "연속 음절 핵"으로 선정한다. 훈련 음성 신호의 전사 정보로부터 획득한 음소를 표 1을 이용하여 분류하고, 분류 항목의 번호는 음소에 해당하는 음성 구간에서

추출된 특징의 출력이 된다. 그런데 보통 훈련 음성 신호의 전사 정보는 음소에 해당하는 시간 정보를 가지고 있지 아니하므로, 음소가 해당하는 음성 신호의 구간을 추정하기 위하여 GMM-HMM(Gaussian Mixture Model-Hidden Markov Model) 기반의 음성 인식기로 강제 정렬(forced alignment)을 수행하여, 음소에 해당하는 음성 신호 구간을 추정한다. 이때, 강제 정렬의 정확도가 높을수록 성능도 향상된다. 잡음 환경에서 획득한 훈련 음성 신호에서는 상태 강제 정렬 정확도가 떨어지므로, 잡음 처리 과정을 거친 후, GMM-HMM 기반의 음성 인식기로 강제 정렬을 수행한다. 이때, 출력이 "음절 핵"에서 "연속 음절 핵"으로 변하는 전이 구간에서 훈련이 잘 되도록, 이웃 프레임들의 출력을 묶어 다중 프레임 출력을 사용할 수 있다. 음절 핵 검출 모델 훈련부는 상기 크기 특징과 위상 특징을 입력으로 하고, 상기 음소 분류 항목의 번호를 출력으로 하여 심층 신경망 모델을 적용한다.

[0063] 음절 핵 검출 모델 훈련부는 음성 핵 검출을 위한 심층 신경망 모델이 적용된 크로스 엔트로피(CE, Cross Entropy)를 기준치로 하여 역전파(back-propagation) 알고리즘을 적용하여 훈련한다.

[0064] 도 4는 인공 신경망의 예측 방법을 설명하기 위한 예시도를 나타낸다.

[0065] 인공 신경망은 최초의 입력 데이터로 이루어진 입력층과 최후의 출력 데이터로 이루어진 출력층을 포함하고, 입력 데이터로부터 출력 데이터를 산출하는 중간 층으로서 은닉층을 포함한다. 은닉층은 하나 이상 존재하며, 2 이상의 은닉층을 포함하는 인공 신경망을 심층 신경망이라 한다. 각 층에 존재하는 노드에서 실제 연산이 이루어지고, 각 노드는 연결선으로 연결된 다른 노드의 출력값을 토대로 연산할 수 있다.

[0066] 도 4에서 보이는 바와 같이 원칙적으로 입력데이터 상호간 또는 동일 층에 속하는 노드들 간에는 서로 영향을 주지 아니하며, 각 층은 상위 또는 하위의 인접한 층의 노드에만 입력값 또는 출력값으로서 서로 데이터를 주고받는 것이다.

[0067] 도 4에서는 층간의 모든 노드 사이에 연결선이 연결되어 있으나, 필요에 따라 인접한 각 층에 속하는 노드 사이에 연결선이 없을 수도 있다. 다만, 연결선이 없는 경우는 해당 입력값에 대하여 가중치를 0으로 설정하여 처리할 수도 있다.

[0068] 인공 신경망의 예측 방향에 따라 입력층으로부터 출력층의 결과값을 예측한 경우, 학습과정에서 결과값들로부터 입력값을 예측할 수 있게 된다. 통상 인공 신경망에 있어서 입력값과 출력값이 일대일 대응관계에 있지 아니하므로, 출력층으로서 입력층을 그대로 복구하는 것은 불가능하나, 예측 알고리즘을 고려하여 역전파(back-propagation, backpropa) 알고리즘에 의해 결과값으로부터 산출된 입력데이터가 최초의 입력데이터와 상이하다면, 인공 신경망의 예측이 부정확하다고 볼 수 있으므로, 제약조건 하에서 산출된 입력 데이터가 최초의 입력 데이터와 유사해지도록 예측 계수를 변경하여 학습을 훈련할 수 있게 된다.

[0069] 도 5는 본 발명의 일실시예에 따른 자연어 인식 성능 개선 방법을 나타내는 흐름도이다.

[0070] 도 5에 따르면, 자연어 인식 개선 방법은 (1) 소정의 시간 간격으로 음성 신호를 복수의 프레임으로 분할하고 이산 푸리에 변환을 적용하여 시간 영역에서 주파수 영역으로 변환하는 단계; (2) 상기 주파수 영역으로 변환된 음성 신호의 크기로부터 크기 특징을 추출하는 단계; (3) 상기 주파수 영역으로 변환된 음성 신호의 위상으로부터 위상 특징을 추출하는 단계; (4) 상기 크기 특징과 위상 특징을 심층 신경망의 입력으로 사용하여 음절 핵을 검출하는 단계; (5) 깨끗한 음성과 잡음의 DFT 계수 분포를 정규 분포로 모델링하고 우도비 검정을 수행함으로써 음성 구간을 검출하는 단계; (6) 상기 검출된 음절 구간의 시간을 이용하여 발성 속도를 결정하는 단계; (7) 상기 발성 속도를 이용하여 시간축 변환 정도를 계산하는 단계; 및 (8) 상기 시간축 변환 정도를 이용하여 음향 모델에 적합하게 음성의 길이를 변환하는 단계;를 포함한다.

[0071] 상기 크기 특징은 멜-스케일 필터뱅크 로그 에너지, MFCC, LPC, 피치, 하모닉 성분, 스펙트럼 평탄도 중 적어도 하나를 포함하는 것이다.

[0072] 상기 위상 특징은 델타-위상 스펙트럼, 위상 왜곡 편차, 그룹 지연, 순환 분산 중 적어도 하나를 포함하는 것이다.

[0073] 상기 시간축 변환 정도는 변화율 또는 중첩률 중 어느 하나인 것이다.

[0074] 도 6은 본 발명의 일실시예에 따른 훈련 음성 신호로부터 심층 신경망을 훈련하는 방법을 나타내는 흐름도이다.

[0075] 도 6에 따르면 심층 신경망 훈련 방법은 (1) 훈련 음성 신호를 주파수 영역으로 변환하는 단계; (2) 상기 변환된 훈련 음성 신호로부터 크기 특징 및 위상 특징을 추출하는 단계; (3) 훈련 음성 신호의 전사 정보에서 획득한 음소를 묵음, 자음, 음절 핵 및 연속 음절 핵 중 어느 하나로 분류하여 심층 신경망에 사용할 출력을 생성하는 단계; 및 (4) 상기 크기 특징과 위상 특징을 입력으로 하여 상기 음소 분류 항목을 출력으로 하는 심층 신경망을 CE를 기준치로 역전파 알고리즘으로 훈련하는 단계;를 포함한다.

[0076] 바람직하게는 상기 (3) 단계는 훈련 음성 신호의 전사 정보가 음소에 해당하는 시간 정보를 가지고 있지 않은 경우, GMM-HMM 기반의 음성 인식기로 강제 정렬을 수행하여 음소에 해당하는 음성 신호 구간을 추정하는 단계를 더 포함할 수 있다.

[0077] 바람직하게는 상기 (3) 단계는 출력이 음절 핵에서 연속 음절 핵으로 변하는 전이 구간에서 훈련이 잘 되도록, 이웃 프레임들의 출력을 묶어 다중 프레임 출력을 사용하는 단계를 더 포함할 수 있다.

[0078] 바람직하게는 출력으로 사용되는 상기 묵음, 자음, 음절 핵, 연속 음절 핵에 대하여 일련 번호(예컨대, 1, 2, 3, 4)를 부여할 수 있다.

[0079] 이상, 본 발명의 구성에 대하여 첨부 도면을 참조하여 상세히 설명하였으나, 이는 예시에 불과한 것으로서, 본 발명이 속하는 기술 분야에 통상의 지식을 가진 자라면 본 발명의 기술적 사상의 범위 내에서 다양한 변형과 변경이 가능함은 물론이다. 따라서 본 발명의 보호 범위는 전술한 실시예에 국한되어서는 아니 되며 이하의 특허청구범위의 기재에 의하여 정해져야 할 것이다.

부호의 설명

[0080] 100: 컴퓨터 시스템

110: 프로세서

120: 메모리

123: ROM

126: RAM

130: 데이터 통신 버스

140: 저장소

150: 사용자 입력 장치

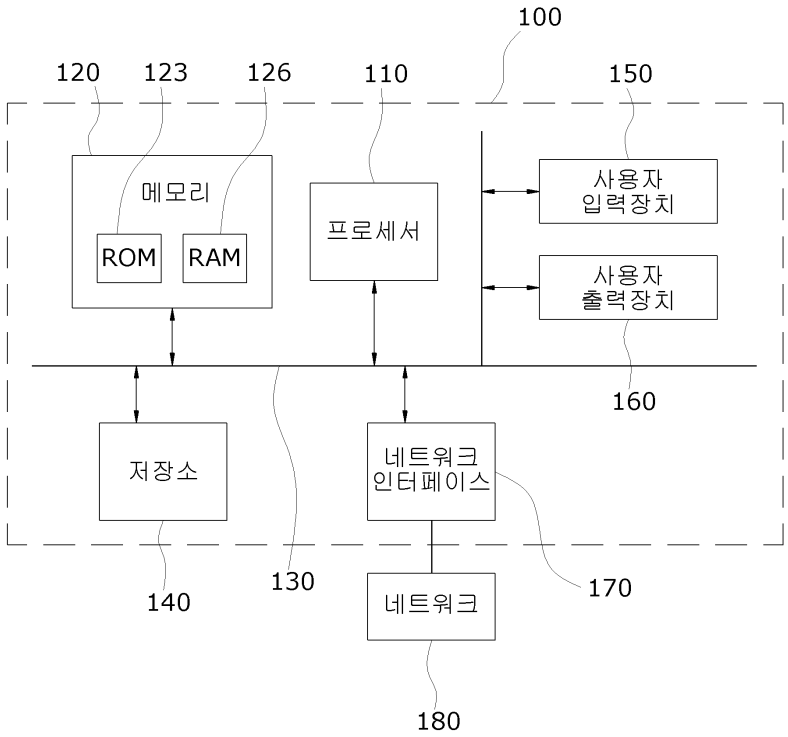
160: 사용자 출력 장치

170: 네트워크 인터페이스

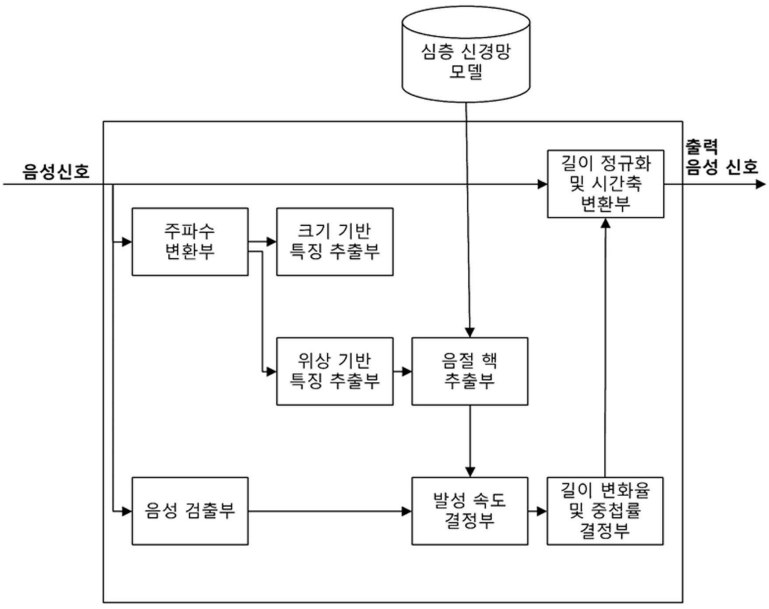
180: 네트워크

도면

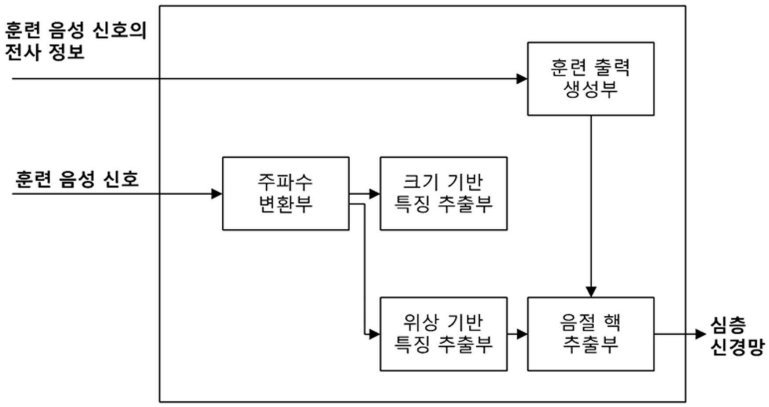
도면1


100
120
123
126
110
150
메모리
ROM
RAM
프로세서
사용자
입력장치
사용자
출력장치
160
저장소
네트워크
인터페이스
140
130
170
네트워크
180

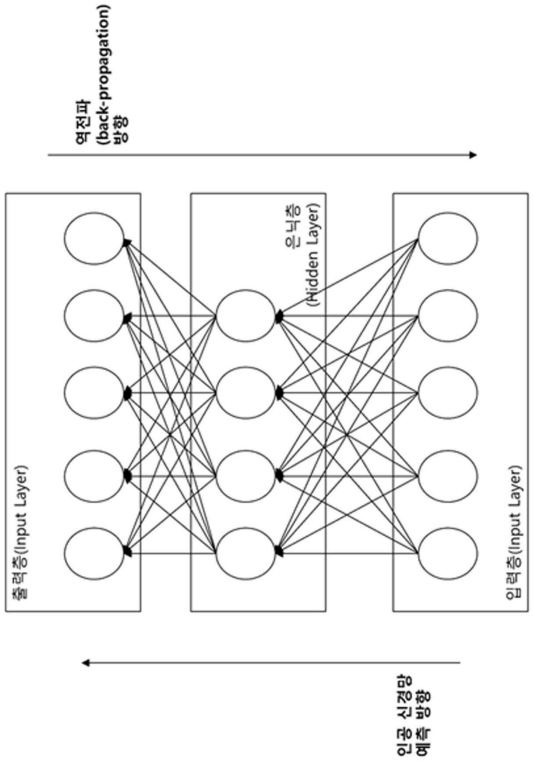
도면2


심층 신경망
모델
음성신호
길이 정규화
및 시간축
변환부
출력
음성 신호
주파수
변환부
크기 기반
특징 추출부
위상 기반
특징 추출부
음절 핵
추출부
음성 검출부
발성 속도
결정부
길이 변화율
및 중첩률
결정부

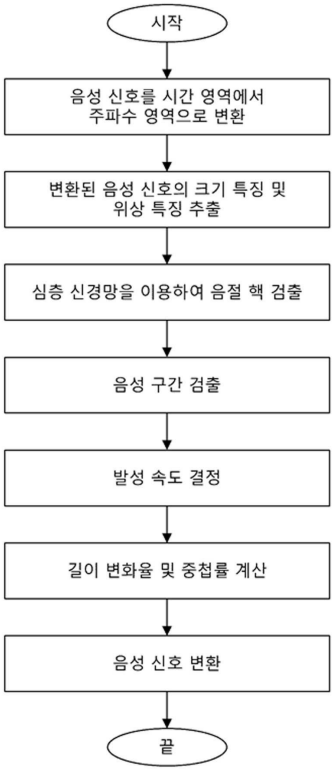
도면3


훈련 음성 신호의
전사 정보
훈련 출력
생성부
훈련 음성 신호
주파수
변환부
크기 기반
특징 추출부
위상 기반
특징 추출부
음절 핵
추출부
심층
신경망

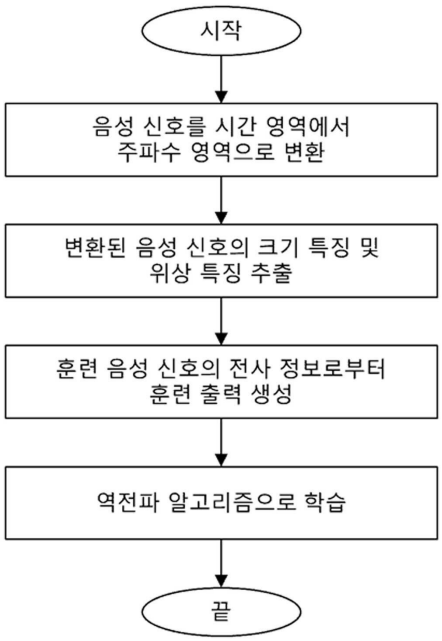
도면4


역전파
(back-propagation)
방향
출력층(Input Layer)
은닉층
(Hidden Layer)
입력층(Input Layer)
인공 신경망
동작 방향

도면5


시작
음성 신호를 시간 영역에서
주파수 영역으로 변환
변환된 음성 신호의 크기 특징 및
위상 특징 추출
심층 신경망을 이용하여 음절 핵 검출
음성 구간 검출
발성 속도 결정
길이 변화율 및 중첩률 계산
음성 신호 변환
끝

도면6


시작
음성 신호를 시간 영역에서
주파수 영역으로 변환
변환된 음성 신호의 크기 특징 및
위상 특징 추출
훈련 음성 신호의 전사 정보로부터
훈련 출력 생성
역전파 알고리즘으로 학습
끝