(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0098083
(43) 공개일자 2021년08월10일

(51) 국제특허분류(Int. Cl.)
G10L 25/48 (2013.01) *G06N 3/04* (2006.01)
G10L 15/16 (2006.01) *G10L 19/038* (2013.01)
G10L 25/24 (2013.01) *G10L 25/93* (2013.01)
(52) CPC특허분류
G10L 25/48 (2013.01)
G06N 3/0454 (2013.01)
(21) 출원번호 10-2020-0011825
(22) 출원일자 2020년01월31일
심사청구일자 2020년01월31일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
강홍구
서울특별시 서대문구 연세로 50, 제3공학관 C523호(신촌동, 연세대학교)
한혜원
서울특별시 서대문구 연세로 50, 제3공학관 C505호(신촌동, 연세대학교)
(뒷면에 계속)
(74) 대리인
특허법인우인

전체 청구항 수 : 총 14 항

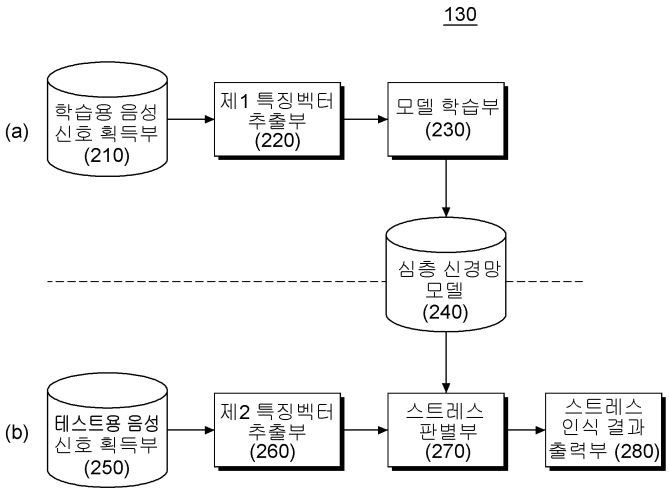
(54) 발명의 명칭 **가중치를 이용한 음성 신호의 스트레스 판별 방법 및 그를 위한 장치**

(57) 요 약

가중치를 이용한 음성 신호의 스트레스 판별 방법 및 그를 위한 장치를 개시한다.

본 발명의 실시예에 따른 음성 신호의 스트레스 판별 방법은, 기 생성된 음성 신호를 획득하는 음성 신호 획득 단계; 상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계; 및 상기 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습하고, 학습 결과에 근거하여 음성 신호의 스트레스가 판별되도록 하는 모델 학습 단계를 포함할 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G10L 15/16 (2013.01)
G10L 19/038 (2013.01)
G10L 25/24 (2013.01)
G10L 25/93 (2013.01)
(72) 발명자
변경근
서울특별시 서대문구 연세로 50, 제3공학관 C505호(신촌동, 연세대학교)
신현경
서울특별시 서대문구 연세로 50, 제3공학관 C505호(신촌동, 연세대학교)

이 발명을 지원한 국가연구개발사업

과제고유번호	2016-0-00562
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송연구개발사업
연구과제명	상대방의 감성을 추론, 판단하여 그에 맞추어 대화하고 대응할 수 있는 감성 지능 연구개발 (4/5)
기 여 율	1/1
과제수행기관명	한국과학기술원
연구기간	2019.05.01 ~ 2020.02.29

명 세 서

청구범위

청구항 1

하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 컴퓨팅 디바이스에 의해 수행되는 스트레스 판별 방법에 있어서, 상기 컴퓨팅 디바이스는,

기 생성된 음성 신호를 획득하는 음성 신호 획득 단계;

상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계; 및

상기 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습하고, 학습 결과에 근거하여 음성 신호의 스트레스가 판별되도록 하는 모델 학습 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 2

제1항에 있어서,

상기 특징 벡터 추출 단계는,

소정의 윈도우 단위로 시간적 흐름에 따라 상기 음성 신호의 주파수 변화를 분석하기 위하여 시간 축에서 음성 신호를 일정한 길이의 프레임 단위로 변환하는 푸리에 변환 단계;

구분된 프레임 각각에 다수의 주파수 영역에 대하여 패턴을 가지고 있는 멜-필터 뱅크를 곱하여 멜 스케일의 각 주파수 대역에 대한 에너지를 나타내는 멜-스펙트로그램을 획득하여 멜-필터 뱅크의 특징을 추출하는 필터 뱅크 처리 단계;

상기 멜-필터 뱅크의 특징을 정규화 처리하는 정규화 단계; 및

기 설정된 고정된 시간의 길이로 정규화된 특징을 분할하여 상기 특징 벡터를 추출하는 분할 처리 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 3

제2항에 있어서,

상기 특징 벡터 추출 단계는,

상기 음성 신호에서 노이즈를 제거하는 노이즈 제거 단계;

프리 엠퍼시스(pre-emphasis) 필터를 통해 고주파수를 강조하여 왜곡을 보상하는 왜곡 보상 단계;

왜곡이 보상된 음성 신호에서 음성이 존재하는 구간을 구분하여 음성 세그먼트를 획득하여 상기 푸리에 변환 단계로 전달하는 묵음 처리 단계

를 추가로 포함하는 것을 음성 신호의 스트레스 판별 방법.

청구항 4

제2항에 있어서,

상기 필터 뱅크 처리 단계는,

주파수 대역간 에너지 스케일 차이를 좁히기 위해 로그 함수를 적용하여 로그 스케일의 에너지로 변환하여 상기 멜-필터 뱅크의 특징을 추출하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 5

제1항에 있어서,

상기 모델 학습 단계는,

상기 특징 벡터를 입력 받고, 상기 특징 벡터의 시간적, 주파수 성분을 스트레스 판별에 적합하도록 모델링하여 프레임별로 출력 벡터를 출력하는 인코더 처리 단계;

프레임별로 출력된 상기 출력 벡터를 기반으로 시간별 가중치 벡터를 계산하고, 상기 시간별 가중치 벡터를 상기 출력 벡터와 연산 처리하여 프레임 레벨의 출력 벡터를 문장 레벨의 벡터로 변환하는 가중치 처리 단계; 및

문장 레벨 특징 벡터와 스트레스 레이블 간의 비선형적 관계를 모델링하여 스트레스의 존재 여부에 대한 판별 결과를 생성하는 분류 처리 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 6

제5항에 있어서,

상기 모델 학습 단계는,

상기 특징 벡터를 기반으로 도출된 상기 판별 결과와 스트레스 레이블 간의 에러를 손실함수로 계산한 뒤 역전파 알고리즘을 이용하여 스트레스 레이블의 에러가 최소가 되도록 학습을 반복하여 수행하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 7

제5항에 있어서,

상기 인코더 처리 단계는,

합성곱 신경망(CNN)과 순환 신경망(RNN)이 결합된 CRNN(Convolutional Recurrent Neural Networks)를 기반으로 동작하며,

상기 특징 벡터는 상기 합성곱 신경망을 통과하여 시간 및 주파수 측면에서 크기가 일정 비율로 감소한 형태의 출력 벡터로 출력되고, 순환 신경망(RNN)은 상기 출력 벡터를 기반으로 프레임별 출력 벡터를 생성하여 출력하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 8

제5항에 있어서,

상기 가중치 처리 단계는,

상기 출력 벡터를 전결합 레이어(fully-connected layer)에 통과시켜 시간별 가중치 벡터를 산출하는 가중치 생성 단계; 및

가중합 풀링(attention pooling)을 이용하여 상기 시간별 가중치 벡터 각각을 상기 출력 벡터에 적용하여 상기 문장 레벨의 벡터로 변환하는 가중치 적용 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 9

제8항에 있어서,

상기 가중치 적용 단계는,

상기 출력 벡터와 상기 시간별 가중치 벡터 각각을 곱(element-wise) 한 후 더하여 상기 문장 레벨의 벡터로 변환하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 10

하나 이상의 프로세서 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는

컴퓨팅 디바이스에 의해 수행되는 스트레스 판별 방법에 있어서, 상기 컴퓨팅 디바이스는,

음성 신호를 획득하는 음성 신호 획득 단계;

상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계;

시간별 구간에 서로 다른 가중치를 부여하여 기 학습된 심층 신경망 모델을 기반으로 상기 특징 벡터에 대한 스트레스의 존재 여부를 판별하는 스트레스 판별 단계; 및

상기 판별 결과에 대한 스트레스 레이블 정보를 출력하는 인식 결과 출력 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 11

제10항에 있어서,

상기 스트레스 판별 단계는,

기 생성된 음성 신호를 획득하는 음성 신호 획득 단계;

상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계; 및

상기 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습하고, 학습 결과에 근거하여 음성 신호의 스트레스가 판별되도록 하는 모델 학습 단계를 포함하여 학습된 상기 심층 신경망 모델을 기반으로 상기 특징 벡터에 대한 스트레스의 존재 여부를 판별하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 12

제11항에 있어서,

상기 모델 학습 단계는,

상기 특징 벡터를 입력 받고, 상기 특징 벡터의 시간적, 주파수 성분을 스트레스 판별에 적합하도록 모델링하여 프레임별로 출력 벡터를 출력하는 인코더 처리 단계;

프레임별로 출력된 상기 출력 벡터를 기반으로 시간별 가중치 벡터를 계산하고, 상기 시간별 가중치 벡터를 상기 출력 벡터와 연산 처리하여 프레임 레벨의 출력 벡터를 문장 레벨의 벡터로 변환하는 가중치 처리 단계; 및

문장 레벨 특징 벡터와 스트레스 레이블 간의 비선형적 관계를 모델링하여 스트레스의 존재 여부에 대한 판별 결과를 생성하는 분류 처리 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 13

제12항에 있어서,

상기 가중치 처리 단계는,

상기 출력 벡터를 전결합 레이어(fully-connected layer)에 통과시켜 시간별 가중치 벡터를 산출하는 가중치 생성 단계; 및

가중합 풀링(attention pooling)을 이용하여 상기 시간별 가중치 벡터 각각을 상기 출력 벡터에 적용하여 상기 문장 레벨의 벡터로 변환하는 가중치 적용 단계

를 포함하는 것을 특징으로 하는 음성 신호의 스트레스 판별 방법.

청구항 14

음성 신호의 스트레스를 판별하는 장치로서,

하나 이상의 프로세서; 및

상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로그램들은 하나

이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서들에서,

음성 신호를 획득하는 음성 신호 획득 단계;

상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계;

시간별 구간에 서로 다른 가중치를 부여하여 기 학습된 심층 신경망 모델을 기반으로 상기 특징 벡터에 대한 스트레스의 존재 여부를 판별하는 스트레스 판별 단계; 및

상기 판별 결과에 대한 스트레스 레이블 정보를 출력하는 인식 결과 출력 단계

를 포함하는 동작들을 수행하게 하는 것을 특징으로 하는 음성 신호의 스트레스 판별 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 시간적 구간에 부여되는 가중치를 이용하여 음성 신호의 스트레스를 판별하는 방법 및 그를 위한 장치에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 음성신호를 이용하여 스트레스를 판별하는 기술은 크게 음성의 특징 벡터를 추출하는 부분과 추출한 벡터와 스트레스의 상태 사이를 통계적 방법으로 모델링하는 부분으로 나뉜다. 기존에 스트레스 판별에 사용하였던 음성의 특징 벡터로는 pitch, MFCC(Mel-frequency cepstral coefficients), 프레임 별 에너지 등이 있다. 이러한 음성 특징벡터들은 기지정된 특징벡터 추출 알고리즘의 과정을 따라 얻을 수 있다.

[0004] 기존의 통계적 모델링 방법으로서는 은닉 마르코프 모델(HMM : Hidden Markov Model), 서포트 벡터 머신(SVM : Support Vector Machine) 등이 있다. 은닉 마르코프 모델은 음성 특징벡터들로부터 마르코프 체인의 성질을 이용하여 음성의 순열적 특성을 모델링한 후, 확률적 계산 알고리즘을 활용하여 데이터를 분류하는 방식이다. SVM은 데이터의 통계적 특성을 활용하여 두 가지의 클래스를 구분할 수 있는 하이퍼플레인(hyperplane)을 최적화하는 방식으로 데이터를 분류하는 모델이다. 이러한 통계적 모델링 방식은 짧은 시간(10-40 ms) 마다 빠르게 통계적 특성이 변화하는 음성 신호의 특성을 찾아내기가 어려운 문제점을 가진다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 음성 신호에서 특징 벡터를 추출하고, 추출된 특징 벡터를 기반으로 시간적 구간에 가중치를 부여하여 심층 신경망 모델을 학습하며, 학습 결과에 따른 심층 신경망 모델을 이용하여 음성 신호의 스트레스를 판별하는 가중치를 이용한 음성 신호의 스트레스 판별 방법 및 그를 위한 장치를 제공하는 데 주된 목적이 있다.

과제의 해결 수단

[0006] 본 발명의 일 측면에 의하면, 상기 목적을 달성하기 위한 음성 신호의 스트레스 판별 방법은, 기 생성된 음성 신호를 획득하는 음성 신호 획득 단계; 상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계; 및 상기 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습하고, 학습 결과에 근거하여 음성 신호의 스트레스가 판별되도록 하는 모델 학습 단계를 포함할 수 있다.

[0007] 또한, 본 발명의 다른 측면에 의하면, 상기 목적을 달성하기 위한 음성 신호의 스트레스 판별 방법은, 음성 신호를 획득하는 음성 신호 획득 단계; 상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계; 시간별 구간에 서로 다른 가중치를 부여하여 기 학습된 심층 신경망 모델을 기반으로 상기 특징 벡터에 대한 스트레스의 존재 여부를 판별하는 스트레스 판별 단계; 및 상기 판별 결과에 대한 스트레

스 레이블 정보를 출력하는 인식 결과 출력 단계를 포함할 수 있다.

[0008] 또한, 본 발명의 다른 측면에 의하면, 상기 목적을 달성하기 위한 성 신호의 스트레스 판별 장치는, 하나 이상의 프로세서; 및 상기 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하며, 상기 프로그램들은 하나 이상의 프로세서에 의해 실행될 때, 상기 하나 이상의 프로세서들에서, 음성 신호를 획득하는 음성 신호 획득 단계; 상기 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출하는 특징 벡터 추출 단계; 시간별 구간에 서로 다른 가중치를 부여하여 기 학습된 심층 신경망 모델을 기반으로 상기 특징 벡터에 대한 스트레스의 존재 여부를 판별하는 스트레스 판별 단계; 및 상기 판별 결과에 대한 스트레스 레이블 정보를 출력하는 인식 결과 출력 단계를 포함하는 동작들을 수행하게 할 수 있다.

발명의 효과

[0009] 이상에서 설명한 바와 같이, 본 발명은 음성 신호를 이용한 사용자 스트레스 판별 모델은 CRNN 기반 네트워크를 사용하여 음성신호에서 입력된 특징벡터로부터 멜-스펙트로그램의 시간적, 주파수적 특성을 효과적으로 모델링 할 수 있다는 효과가 있다.

[0010] 또한, 본 발명은 음성신호에서 스트레스를 효과적으로 표현하는 구간의 가중치를 어텐션(attention) 알고리즘으로 자동적으로 찾아 가중치를 부여함으로써 스트레스 판별 성능을 향상시킬 수 있는 효과가 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 실시예에 따른 스트레스 판별 장치를 개략적으로 나타낸 블록 구성도이다.

도 2는 본 발명의 실시예에 따른 프로세서의 동작 구성을 개략적으로 나타낸 블록 구성도이다.

도 3a 및 도 3b는 본 발명의 실시예에 따른 스트레스 판별 방법을 설명하기 위한 순서도이다.

도 4는 본 발명의 실시예에 따른 음성 신호에서 특징 벡터를 추출하는 동작 구성을 나타낸 블록 구성도이다.

도 5는 본 발명의 실시예에 따른 모델 학습의 동작 구성을 나타낸 블록 구성도이다.

도 6은 본 발명의 실시예에 따른 심층 신경망 모델의 네트워크 구조를 나타낸 도면이다.

도 7은 본 발명의 실시예에 따른 심층 신경망 모델의 스트레스 판별 동작을 설명하기 위한 예시도이다.

도 8은 본 발명의 실시예에 따른 스트레스 판별 결과를 나타낸 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하, 본 발명의 바람직한 실시예를 첨부된 도면들을 참조하여 상세히 설명한다. 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다. 또한, 이하에서 본 발명의 바람직한 실시예를 설명할 것이나, 본 발명의 기술적 사상은 이에 한정하거나 제한되지 않고 당업자에 의해 변형되어 다양하게 실시될 수 있음은 물론이다. 이하에서는 도면들을 참조하여 본 발명에서 제안하는 가중치를 이용한 음성 신호의 스트레스 판별 방법 및 그를 위한 장치에 대해 자세하게 설명하기로 한다.

[0013] 도 1은 본 발명의 실시예에 따른 스트레스 판별 장치를 개략적으로 나타낸 블록 구성도이다.

[0014] 본 실시예에 따른 스트레스 판별 장치(100)는 입력부(110), 출력부(120), 프로세서(130), 메모리(140) 및 데이터 베이스(150)를 포함한다. 도 1의 스트레스 판별 장치(100)는 일 실시예에 따른 것으로서, 도 1에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 스트레스 판별 장치(100)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 스트레스 판별 장치(100)는 컴퓨팅 디바이스로 구현될 수 있고, 스트레스 판별 장치(100)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 장치로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.

[0015] 스트레스 판별 장치(100)는 음성 신호에서 특징 벡터를 추출하고, 추출된 특징 벡터를 기반으로 시간적 구간에 가중치를 부여하여 심층 신경망 모델을 학습하며, 학습 결과에 따른 심층 신경망 모델을 이용하여 음성 신호의 스트레스를 판별하는 동작을 수행한다.

[0016] 입력부(110)는 스트레스 판별 장치(100)에서의 영상 생성 동작을 수행하기 위한 신호 또는 데이터를 입력하거나 획득하는 수단을 의미한다. 입력부(110)는 프로세서(130)와 연동하여 다양한 형태의 신호 또는 데이터를 입력하

거나, 외부 장치와의 연동을 통해 신호 또는 데이터를 획득하여 프로세서(130)로 전달할 수도 있다. 여기서, 입력부(110)는 음성 신호, 특징 벡터 등을 입력하기 위한 모듈로 구현될 수 있으나 반드시 이에 한정되는 것은 아니다.

[0017] 출력부(120)는 프로세서(130)와 연동하여 특징 벡터, 스트레스 판별 결과, 심층 신경망 모델의 학습 결과 등 다양한 정보를 출력할 수 있다. 출력부(120)는 스트레스 판별 장치(100)에 구비된 디스플레이(미도시)를 통해 다양한 정보를 출력할 수 있으나 반드시 이에 한정되는 것은 아니며, 다양한 형태의 방식으로 출력을 수행할 수 있다.

[0018] 프로세서(130)는 메모리(140)에 포함된 적어도 하나의 명령어 또는 프로그램을 실행시키는 기능을 수행한다.

[0019] 본 실시예에 따른 프로세서(130)는 입력부(110) 또는 데이터 베이스(150)로부터 음성 신호를 획득하고, 음성 신호에 대한 특징 벡터를 추출하는 동작을 수행한다.

[0020] 또한, 프로세서(130)는 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습하고, 학습 결과를 기반으로 학습된 심층 신경망 모델을 생성 및 저장하고, 음성 신호의 스트레스 판별시 심층 신경망 모델을 적용하여 스트레스 판별을 수행한다.

[0021] 메모리(140)는 프로세서(130)에 의해 실행 가능한 적어도 하나의 명령어 또는 프로그램을 포함한다. 메모리(140)는 음성 신호, 특징 벡터, 스트레스 판별 결과, 학습 결과 등을 처리하는 동작을 위한 명령어 또는 프로그램을 포함할 수 있다.

[0022] 데이터베이스(150)는 데이터베이스 관리 프로그램(DBMS)을 이용하여 컴퓨터 시스템의 저장공간(하드디스크 또는 메모리)에 구현된 일반적인 데이터구조를 의미하는 것으로, 데이터의 검색(추출), 삭제, 편집, 추가 등을 자유롭게 행할 수 있는 데이터 저장형태를 뜻하는 것으로, 오라클(Oracle), 인포믹스(Infomix), 사이베이스(Sybase), DB2와 같은 관계형 데이타베이스 관리 시스템(RDBMS)이나, 겜스톤(Gemston), 오리온(Orion), O2 등과 같은 객체 지향 데이타베이스 관리 시스템(OODBMS) 및 엑셀론(Excelon), 타미노(Tamino), 세카이주(Sekaiju) 등의 XML 전용 데이터베이스(XML Native Database)를 이용하여 본 발명의 일 실시예의 목적에 맞게 구현될 수 있고, 자신의 기능을 달성하기 위하여 적당한 필드(Field) 또는 엘리먼트들을 가지고 있다.

[0023] 본 실시예에 따른 데이터베이스(150)는 음성 신호의 스트레스 판별을 위한 학습과 관련된 데이터를 저장하고, 기 저장된 음성 신호의 스트레스 판별을 위한 학습과 관련된 데이터를 제공할 수 있다.

[0024] 데이터베이스(150)에 저장된 데이터는 음성 신호, 특징 벡터, 스트레스 판별 결과, 학습 결과 등에 대한 데이터일 수 있다. 데이터베이스(140)는 스트레스 판별 장치(100) 내에 구현되는 것으로 기재하고 있으나 반드시 이에 한정되는 것은 아니며, 별도의 데이터 저장장치로 구현될 수도 있다.

[0025] 도 2는 본 발명의 실시예에 따른 프로세서의 동작 구성을 개략적으로 나타낸 블록 구성도이다.

[0026] 본 실시예에 따른 스트레스 판별 장치(100)에 포함된 프로세서(130)는 딥 러닝 학습을 기반으로 음성 신호로부터 스트레스를 판별하는 동작을 수행한다. 여기서, 딥 러닝 학습은 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), CRNN(Convolution Recurrent Neural Network) 등을 이용한 학습일 수 있으나 반드시 이에 한정되는 것은 아니다.

[0027] 스트레스 판별 장치(100)에 포함된 프로세서(130)는 음성 신호에서 특징 벡터를 추출하고, 추출된 특징 벡터를 기반으로 시간적 구간에 가중치를 부여하여 심층 신경망 모델을 학습하며, 학습 결과에 따른 심층 신경망 모델을 이용하여 음성 신호의 스트레스를 판별하는 동작이 수행되도록 하며, 음성 신호의 스트레스를 인식 또는 판별하는 모든 기기에 탑재되거나, 스트레스 인식 또는 판별을 수행하는 소프트웨어와 연동할 수 있다.

[0028] 본 실시예에 따른 프로세서(130)는 학습용 음성 신호 획득부(210), 제1 특징벡터 추출부(220), 모델 학습부(230), 심층 신경망 모델 처리부(240), 테스트용 음성 신호 획득부(250), 제2 특징벡터 추출부(260), 스트레스 판별부(270), 스트레스 인식 결과 출력부(280)를 포함할 수 있다. 도 2의 프로세서(130)는 일 실시예에 따른 것으로서, 도 2에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 프로세서(130)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 프로세서(130)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 장치로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.

[0029] 학습용 음성 신호 획득부(210)는 심층 신경망 모델의 학습을 위하여 음성 신호를 데이터베이스로부터 획득한다. 여기서, 음성 신호는 신경 신경망 모델의 학습을 위하여 기 녹음된 음성 신호일 수 있으나 반드시 이에 한정되

는 것은 아니다.

[0030] 제1 특징벡터 추출부(220)는 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출한다. 제1 특징벡터 추출부(220)에서 특징 벡터를 추출하는 동작은 도 4에서 자세히 설명하도록 한다.

[0031] 모델 학습부(230)는 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습하여 심층 신경망 모델을 생성하는 동작을 수행한다. 모델 학습부(230)에서 심층 신경망 모델을 학습하는 동작을 도 5에서 자세히 설명하도록 한다.

[0032] 심층 신경망 모델 처리부(240)는 모델 학습부(230)에서 생성된 심층 신경망 모델을 저장하고, 음성 신호의 스트레스 판별시 심층 신경망 모델을 스트레스 판별부(270)로 제공한다.

[0033] 테스트용 음성 신호 획득부(250)는 스트레스를 판별하기 위한 음성 신호를 데이터베이스로부터 획득한다. 여기서, 음성 신호는 테스트용 음성 신호인 것이 바람직하나 반드시 이에 한정되는 것은 아니며, 실시간으로 수집된 음성 신호일 수도 있다.

[0034] 제2 특징벡터 추출부(260)는 획득된 음성 신호를 소정의 윈도우 단위로 분석하여 특징 벡터를 추출한다. 제2 특징벡터 추출부(260)에서 특징 벡터를 추출하는 방식은 제1 특징벡터 추출부(220)에서 특징 벡터를 추출하는 방식과 동일함에 따라 중복되는 기재는 생략하도록 한다.

[0035] 제2 특징벡터 추출부(260)는 음성 신호에서 추출된 특징 벡터를 스트레스 판별부(270)로 전달한다.

[0036] 스트레스 판별부(270)는 기 학습된 심층 신경망 모델에 추출된 특징 벡터를 입력하여 스트레스의 존재 여부를 판별한다. 스트레스 판별부(270)는 음성 신호에 스트레스를 포함하고 있는지 여부에 대한 스트레스 레이블 정보를 생성한다.

[0037] 스트레스 인식 결과 출력부(280)는 스트레스 판별부(270)에서 생성된 스트레스 레이블 정보를 출력한다.

[0038] 본 발명의 다른 실시예에 따른 스트레스 판별 장치(100)에 포함된 프로세서(130)는 음성 데이터베이스로부터 획득된 음성 신호에서 특징 벡터를 추출하는 특징 벡터 획득부(220), 추출된 특징 벡터를 입력으로 받아 심층 신경망 모델 내부의 파라미터를 업데이트하는 모델 학습부(230), 음성 신호의 스트레스의 유/무 상태를 판단하는 스트레스 판별부(270)로 구성될 수 있다.

[0039] 특징 벡터 획득부(220)는 기 지정된 방식으로 음성 신호의 특성을 반영한 벡터를 추출하는 동작을 수행한다. 시간 축에서 음성 신호를 일정한 길이(5~40ms)의 프레임 단위로 구분한 후, 각 프레임에 대하여 주파수 대역별 에너지를 추출한다. 예를 들어, 특징 벡터 획득부(220)는 음성 신호의 시간, 주파수에 따른 에너지를 나타내는 파워 스펙트로그램을 추출할 수 있다.

[0040] 이후, 특징 벡터 획득부(220)는 다수의 주파수 영역에 대하여 패턴을 가지고 있는 멜-필터 뱅크를 곱하여 멜 스케일의 각 주파수 대역에 대한 에너지를 나타내는 멜-스펙트로그램을 획득한다. 이 때, 특징 벡터 획득부(220)는 주파수 대역간 에너지 스케일 차이를 좁히기 위해 로그 함수를 적용하여 로그 스케일의 에너지로 변환하여 특징 벡터를 추출한다.

[0041] 모델 학습부(230)는 초기 심층 신경망의 초기 모델을 역전파 알고리즘을 통해 파라미터값이 데이터의 통계적 특성을 반영하도록 학습하는 동작을 수행한다.

[0042] 모델 학습부(230)는 입력단에서 추출된 음성 특징 벡터를 입력으로 받고, 출력단에서 음성 신호의 스트레스/비스트레스를 나타내는 레이블 정보를 출력한다.

[0043] 모델 학습부(230)는 입력으로부터 네트워크가 예측한 값과 레이블 간의 에러를 손실함수로 계산한 뒤 역전파 알고리즘을 이용하여 레이블의 에러가 최소가 되도록 학습한다.

[0044] 스트레스 판별부(270)는 학습된 심층 신경망 모델에 신규로 획득되거나 테스트용으로 획득된 음성 신호의 음성 특징 벡터를 통과시켜 주어진 음성 신호의 스트레스 유/무를 판별한다.

[0045] 도 3a 및 도 3b는 본 발명의 실시예에 따른 스트레스 판별 방법을 설명하기 위한 순서도이다.

[0046] 도 3a는 스트레스 판별 장치(100)에서 심층 신경망 모델의 학습(훈련)을 수행하기 위한 스트레스 판별 방법에 대한 순서도이고, 도 3b는 스트레스 판별 장치(100)에서 기 학습된 심층 신경망 모델을 적용하여 음성 신호의 스트레스를 판별하는 동작을 수행하기 위한 스트레스 판별 방법에 대한 순서도이다.

[0047] 도 3a를 참고하면, 스트레스 판별 장치(100)는 심층 신경망 모델의 학습을 위하여 음성 신호를 데이터베이스로부터 획득한다(S310). 여기서, 음성 신호는 신경 신경망 모델의 학습을 위하여 기 녹음된 음성 신호일 수 있으나 반드시 이에 한정되는 것은 아니다.

[0048] 스트레스 판별 장치(100)는 음성 신호를 소정의 윈도우 단위로 분석하여 제1 특징 벡터를 추출한다(S320).

[0049] 스트레스 판별 장치(100)는 제1 특징 벡터를 입력으로 프레임별 특징 벡터를 모델링하고, 가중치를 부여하여 심층 신경망 모델을 학습한다(S330).

[0050] 스트레스 판별 장치(100)는 학습된 심층 신경망 모델을 생성 및 저장하고, 음성 신호의 스트레스 판별시 심층 신경망 모델을 스트레스 판별을 위해 제공한다(S340).

[0051] 도 3b를 참고하면, 스트레스 판별 장치(100)는 스트레스를 판별하기 위한 음성 신호를 데이터베이스로부터 획득한다(S350). 여기서, 음성 신호는 테스트용 음성 신호인 것이 바람직하나 반드시 이에 한정되는 것은 아니며, 실시간으로 수집된 음성 신호일 수도 있다.

[0052] 스트레스 판별 장치(100)는 획득된 음성 신호를 소정의 윈도우 단위로 분석하여 제2 특징 벡터를 추출한다(S360).

[0053] 스트레스 판별 장치(100)는 기 학습된 심층 신경망 모델에 추출된 제2 특징 벡터를 입력하여 스트레스의 존재 여부를 판별한다(S370). 스트레스 판별 장치(100)는 음성 신호에 스트레스를 포함하고 있는지 여부에 대한 스트레스 레이블 정보를 생성한다.

[0054] 스트레스 판별 장치(100)는 생성된 스트레스 레이블 정보를 출력한다(S380).

[0055] 도 3a 및 도 3b에서는 각 단계를 순차적으로 실행하는 것으로 기재하고 있으나, 반드시 이에 한정되는 것은 아니다. 다시 말해, 도 3a 및 도 3b에 기재된 단계를 변경하여 실행하거나 하나 이상의 단계를 병렬적으로 실행하는 것으로 적용 가능할 것이므로, 도 3a 및 도 3b는 시계열적인 순서로 한정되는 것은 아니다.

[0056] 도 3a 및 도 3b에 기재된 본 실시예에 따른 스트레스 판별 방법은 애플리케이션(또는 프로그램)으로 구현되고 단말장치(또는 컴퓨터)로 읽을 수 있는 기록매체에 기록될 수 있다. 본 실시예에 따른 스트레스 판별 방법을 구현하기 위한 애플리케이션(또는 프로그램)이 기록되고 단말장치(또는 컴퓨터)가 읽을 수 있는 기록매체는 컴퓨팅 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치 또는 매체를 포함한다.

[0057] 도 4는 본 발명의 실시예에 따른 음성 신호에서 특징 벡터를 추출하는 동작 구성을 나타낸 블록 구성도이다.

[0058] 본 실시예에 따른 제1 특징벡터 추출부(220)는 노이즈 제거부(410), 왜곡 보상부(420), 묵음 처리부(430), 푸리에 변환부(440), 필터 뱅크부(450), 정규화부(460) 및 분할 처리부(470)를 포함한다. 도 4의 제1 특징벡터 추출부(220)는 일 실시예에 따른 것으로서, 도 4에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 제1 특징벡터 추출부(220)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 제2 특징벡터 추출부(260)는 입력되는 음성 신호만 상이할 뿐, 특징 벡터를 추출하는 동작은 제1 특징벡터 추출부(220)와 동일한 구성을 가진다.

[0059] 제1 특징벡터 추출부(220)에서, 심층 신경망 모델의 학습을 위한 강인한 특징을 추출하려면, 녹음된 음성 신호에 복수의 전처리 단계를 적용해야 한다. 이하, 제1 특징벡터 추출부(220)의 전처리 단계에 대해 설명하도록 한다.

[0060] 노이즈 제거부(410)는 음성 신호에서 노이즈를 제거하는 동작을 수행한다. 예를 들어, 노이즈 제거부(410)는 원하지 않는 배경 노이즈의 구성을 제거할 수 있다. 여기서, 노이즈 제거부(410)는 위너 필터링(Wiener filtering)을 적용하여 노이즈를 제거하는 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0061] 왜곡 보상부(420)는 프리 엠퍼시스(pre-emphasis) 필터 즉, 간단한 형태의 고주파 필터(high pass filter)를 적용하여 주파수 스펙트럼의 동적 범위를 줄여 음성신호 처리 과정에서 발생하는 왜곡을 보상한다. 왜곡 보상부(420)는 프리 엠퍼시스(pre-emphasis) 필터를 통해 고주파수를 강조하여 고주파수 영역과 저주파수 영역 사이의 동적 범위의 균형을 맞출 수 있다.

[0062] 묵음 처리부(430)는 음성 신호에서 음성이 존재하는 구간인지 아닌지를 구분하는 동작을 수행한다. 묵음 처리부(430)는 VAD(Voice Activity Detection) 알고리즘을 적용하여 음성 세그먼트를 획득한다. 묵음 처리부(430)는 음성 신호에서 묵음 구간을 탐색하여 제거(0으로 처리)한 후 음성 세그먼트를 획득한다.

[0063] 푸리에 변환부(440)는 소정의 윈도우 단위로 분석 및 변환을 수행한다. 예를 들어, 푸리에 변환부(440)는 입력된 음성 신호를 25 ms의 해닝 윈도우(Hanning window)로 10 ms 마다 분석할 수 있다.

[0064] 푸리에 변환부(440)는 시간적 흐름에 따라 상기 음성 신호의 주파수 변화를 분석하기 위하여 STFT(Short-Time Fourier Transform)를 수행할 수 있다. 다시 말해, 푸리에 변환부(440)는 시간 축에서 음성 신호를 일정한 길이(5 ~ 40 ms)의 프레임 단위로 구분할 수 있다.

[0065] 필터 뱅크부(450)는 구분된 프레임 각각에 대하여 주파수 대역별 에너지를 추출할 수 있다. 예를 들어, 필터 뱅크부(450)는 음성 신호의 시간, 주파수에 따른 에너지를 나타내는 파워 스펙트로그램을 추출할 수 있다.

[0066] 필터 뱅크부(450)는 다수의 주파수 영역에 대하여 패턴을 가지고 있는 멜-필터 뱅크를 곱하여 멜 스케일의 각 주파수 대역에 대한 에너지를 나타내는 멜-스펙트로그램을 획득한다.

[0067] 필터 뱅크부(450)는 주파수 대역간 에너지 스케일 차이를 좁히기 위해 로그 함수를 적용하여 로그 스케일의 에너지로 변환하여 특징을 추출한다.

[0068] 정규화부(460)는 멜-필터 뱅크의 특징을 정규화 처리한다. 정규화부(460)는 멜-필터 뱅크 특징이 제로 평균 및 단위 분산을 갖도록 정규화 처리 한다.

[0069] 분할 처리부(470)는 스트레스 판별 알고리즘의 설정에 따라 고정된 시간의 길이(예: 2 초, 4 초, 5 초 등)로 특징을 분할하여 최종적으로 추출된 특징 벡터를 출력한다.

[0070] 분할 처리부(470)에서 출력된 특징 벡터는 스트레스 판별을 위한 심층 신경망 모델을 학습하기 위하여 모델 학습부(230)로 전달된다.

[0071] 도 5는 본 발명의 실시예에 따른 모델 학습의 동작 구성을 나타낸 블록 구성도이다.

[0072] 본 실시예에 따른 모델 학습부(230)는 음성 특징벡터 획득부(510), 인코더(520), 가중치 생성부(530), 가중치 적용부(532), 분류 처리부(540) 및 결과 출력부(550)를 포함한다. 도 4의 모델 학습부(230)는 일 실시예에 따른 것으로서, 도 4에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 모델 학습부(230)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다.

[0073] 음성 특징벡터 획득부(510)는 음성 신호에서 추출된 음성 특징벡터를 획득한다. 여기서, 음성 특징벡터는 제1 특징벡터 추출부(220)에서 추출된 특징벡터인 것이 바람직하나 반드시 이에 한정되는 것은 아니다.

[0074] 음성 특징벡터 획득부(510)는 학습용 음성 신호를 기반으로 추출된 음성 특징 벡터를 입력 받을 수 있으나, 데이터베이스에 기 저장된 다양한 음성 신호 또는 실시간으로 녹음된 음성 신호에서 추출된 음성 특징벡터를 획득할 수도 있다.

[0075] 인코더(520)는 음성 특징 벡터를 멜-스펙트로그램을 이용하여 스트레스 판별을 하기 적합한 형태의 임베딩 벡터(embedding vector)로 변환하는 동작을 수행한다. 또한, 인코더(520)는 입력된 음성 특징 벡터의 시간적, 주파수 성분을 스트레스 인식에 적합하도록 모델링하는 역할을 수행한다.

[0076] 인코더(520)는 적어도 두 개의 서로 다른 신경망이 결합된 형태로 구성될 수 있다. 구체적으로, 인코더(520)는 합성곱 신경망(CNN)과 순환 신경망(RNN)이 결합된 CRNN(Convolutional Recurrent Neural Networks)를 기반으로 구성된다.

[0077] 합성곱 신경망(CNN)의 입력층에서는 제1 특징벡터 추출부(220)에서 소정의 시간 간격(예: 10 ms)으로 추출되는 멜-스펙트로그램이 입력된다. 합성곱 신경망을 통과하여 시간 및 주파수 측면에서 크기가 일정 비율로 감소한 형태의 출력 벡터를 생성한다. 합성곱 신경망의 출력 벡터는 순환 신경망(RNN)으로 전달되며, 순환 신경망(RNN)은 출력 벡터를 기반으로 프레임별 출력 벡터를 생성하여 가중치 처리부로 전달한다. 여기서, 가중치 처리부(530, 532)는 가중치 생성부(530)과 가중치 적용부(532)로 구성될 수 있다.

[0078] 가중치 생성부(530)는 인코더(520)에서 프레임별로 출력된 출력 벡터 각각에 대한 시간별 가중치를 산출한다.

[0079] 가중치 생성부(530)는 출력 벡터 각각을 전결합 레이어(fully-connected layer)에 통과시켜 시간별 가중치를 산출할 수 있다. 즉, 가중치 생성부(530)는 출력 벡터 각각을 전결합 레이어(fully-connected layer)에 통과시켜 스트레스 레이블 간의 관계를 계산하여 시간별 가중치 벡터를 산출한다. 여기서, 시간별 가중치 벡터는 출력 벡터와 동일한 차원을 가지며, 시간별로 각각 산출된 가중치를 의미한다.

[0080] 가중치 적용부(532)는 인코더(520)의 출력 벡터를 획득하고, 인코더(520)의 출력 벡터와 가중치 생성부(530)에서 산출된 시간별 가중치 벡터 각각을 곱(element-wise) 한 후 더하여 문장 레벨의 벡터로 변환한다. 가중치 적용부(532)는 가중치 벡터를 출력 벡터와 연산 처리하여 프레임 레벨의 벡터를 문장을 대표하는 문장 레벨의 벡터로 변환하고, 변환된 문장 레벨의 벡터는 분류 처리부(540)로 전달된다. 여기서, 가중치 적용부(532)는 가중합 풀링(attention pooling)을 이용하여 시간별 가중치를 출력 벡터에 적용할 수 있다.

[0081] 분류 처리부(540)는 문장 레벨 특징 벡터와 스트레스 레이블 간의 비선형적 관계를 모델링하는 동작을 수행한다. 분류 처리부(540)는 전결합 레이어(fully-connected layer) 및 소프트맥스(Softmax)로 구성될 수 있다.

[0082] 문장 레벨 특징 벡터는 분류 처리부(540)를 통과함으로써, 음성 신호의 스트레스 존재 여부를 판별할 수 있다.

[0083] 결과 출력부(550)는 스트레스의 존재 여부에 대한 판별 결과를 출력한다. 결과 출력부(550)는 판별 결과를 기반으로 심층 신경망 모델이 학습되도록 제공한다.

[0084] 모델 학습부(230)는 스트레스의 존재 여부에 대한 판별 결과와 스트레스 레이블 간의 에러를 계산하여 역전파 알고리즘으로 학습(훈련)된다. 즉, 모델 학습부(230)는 입력된 음성 신호의 음성 특징 벡터를 기반으로 도출된 판별 결과와 스트레스 레이블 간의 에러를 손실함수로 계산한 뒤 역전파 알고리즘을 이용하여 레이블의 에러가 최소가 되도록 학습을 반복하여 수행한다.

[0085] 도 6은 본 발명의 실시예에 따른 심층 신경망 모델의 네트워크 구조를 나타낸 도면이다.

[0086] 본 실시예에 따른 심층 신경망 모델은 모델 학습부(230)에서 학습(훈련)되고, 심층 신경망 모델 처리부(240)에 저장되며, 음성 신호의 스트레스를 판별하기 위한 동작 시 스트레스 판별부(270)에 적용될 수 있다.

[0087] 심층 신경망 모델은 음성 특징 벡터를 입력 받고, 음성 특징 벡터를 멜-스펙트로그램을 이용하여 스트레스 판별을 하기 적합한 형태의 임베딩 벡터(embedding vector)로 변환하는 인코더(encoder), 프레임별로 추출된 임베딩 벡터와 스트레스 레이블 간의 관계를 계산하여 가중치를 부여한 후, 가중합으로 문장 레벨의 음성 특성을 추출하는 어텐션 가중합부, 문장 레벨의 특징 벡터를 이용해 스트레스 판별을 진행하는 문장 레벨 은닉층을 포함하는 형태로 구현될 수 있다.

[0088] 심층 신경망 모델의 네트워크 하단에 있는 인코더(620, 622)는 입력된 음성 특징 벡터의 시간적, 주파수 성분을 스트레스 인식에 적합하도록 모델링하는 역할을 한다.

[0089] 인코더(620, 622)는 여러 층의 합성곱 신경망(620)과 장단기 기억 신경망(LSTM, 622)이 결합한 형태로 구성된다. 네트워크 합성곱 신경망(620)의 입력층으로는 특징 벡터 획득부에서 소정의 시간 간격(예: 10 ms)으로 추출되는 멜-스펙트로그램을 네트워크의 입력으로 사용한다. 합성곱 신경망을 통과하여 시간 및 주파수 측면에서 크기가 일정 비율로 감소한 형태의 출력 벡터를 생성한다.

[0090] 합성곱 신경망(620)의 출력 벡터는 장단기 기억 신경망(622) 네트워크로 전달되며, 장단기 기억 신경망(622)은 출력 벡터를 기반으로 매 프레임별 출력 벡터를 생성하여 어텐션 가중합부(630, 632, 634)으로 전달한다.

[0091] 어텐션 가중합부(630, 632, 634)는 어텐션 은닉층(attention layer, 630)을 포함하며, 어텐션 은닉층(630)은 인코더에서 프레임별로 출력한 벡터 각각에 대하여 스트레스 레이블 간의 관계를 계산하여 시간별 가중치를 산출한다. 본 발명에서 어텐션 은닉층(630)은 각 프레임별 벡터에 대한 전결합 레이어(fully-connected layer)를 추가하여 가중치 벡터(632)가 계산되도록 한다.

[0092] 어텐션 가중합부(630, 632, 634)는 계산된 가중치 벡터와 인코더 출력 값과의 곱을 각각 곱한 후 더하는 가중합 풀링(attention pooling, 634)을 이용하여 프레임 레벨의 벡터를 문장을 대표하는 문장 레벨의 벡터로 변환한다. 이러한 문장 레벨의 벡터는 문장 레벨 은닉층(640, 642)으로 전달된다.

[0093] 문장 레벨 은닉층(640, 642)은 문장 레벨 특징 벡터와 스트레스 레이블 간의 비선형적 관계를 모델링하는 전결합 레이어(640) 및 소프트맥스(642)로 구성되어 있다. 문장 레벨 은닉층(640, 642)을 통과함으로써 스트레스의 유/무를 판별할 수 있다. 전체의 심층 신경망 모델은 네트워크의 출력값과 스트레스 레이블 간의 에러를 계산하여 역전파 알고리즘으로 학습(훈련)된다.

[0094] 도 7은 본 발명의 실시예에 따른 심층 신경망 모델의 스트레스 판별 동작을 설명하기 위한 예시도이다.

[0095] 도 7은 심층 신경망 모델의 스트레스 판별 동작을 위한 CRNN-어텐션 네트워크의 구조를 나타낸다.

[0096] CRNN-어텐션 네트워크는 CNN 처리 단계, RNN 처리 단계, 어텐션 단계, 분류 단계 등으로 구성될 수 있다.

[0097] 도 7을 참조하면, 본 실시예에 따른 심층 신경망 모델은 CNN 처리 단계(720)에서, 음성 신호(710)의 비선형성 특징을 효과적으로 포착하기 위해 최소 단위의 수용 영역 예를 들어, 3 x 3의 필터를 사용할 수 있다. 또한, CNN 처리 단계(720)에서, 레이어 수는 종래 11 개에서 7 개로 줄여 가벼운 모델을 생성하였다.

[0098] CNN 처리 단계(720)에서는, 종래의 VGG-A 모델에서 컨볼루셔널 레이어에서 반복 레이어로 중요한 음향 특징 맵을 유지하면서, 마지막 하나의 컨볼루션 레이어와 최대 풀링 레이어 및 전결합 레이어를 생략한다. 배치 정규화 층은 정류 선형 유닛(ReLU) 활성화 기능 및 최대 풀링 층을 적용하기 전에 이어진다. 컨볼루션 블록(722)의 끝에서 추출된 특징은 주파수 축을 따라 쌓인다.

[0099] RNN 처리 단계(724)는 반복 블록 내에 장단기 기억 신경망(LSTM)으로 구성될 수 있고, 장단기 기억 신경망(LSTM)의 구조는 순차적 음성 신호를 효과적으로 처리하기 위해 사용된다.

[0100] 배치 정규화 계층은 이후 학습 속도를 가속화하고 네트워크를 통한 그라디언트 흐름을 개선함으로써 분류 성능을 향상시키기 위해 이어진다.

[0101] RNN 처리 단계(724)의 반복 블록의 마지막 히든 레이어는 가중합 처리를 위하여 어텐션 단계(730)의 어텐션 블록과 연결된다.

[0102] 가중합 처리 처리된 벡터는 전결합 레이어(740)를 통해 출력되며, 여기서, 출력은 음성 신호가 스트레스 또는 비스트레스 카테고리로 분류되는지를 확인기 위하여 소프트맥스(softmax) 함수를 사용하여 이진 분류기(742)에 연결된다.

[0103] 도 8은 본 발명의 실시예에 따른 스트레스 판별 결과를 나타낸 예시도이다.

[0104] 도 8의 좌측 히트맵은 스트레스가 없는 상태의 음성 신호를 나타내고, 도 8의 우측 히트맵은 스트레스가 존재하는 상태의 음성 신호를 나타낸다.

[0105] 히트맵의 가로 축은 시간(초 단위)을 나타내고, 세로 축은 필터뱅크 번호를 나타낸다.

[0106] 히트맵에서 밝은 색을 나타낼 수록 어텐션의 가중치가 크게 나타남을 보여준다. 도 8을 참조하면, 31 번째 필터뱅크에 대해 전체적으로 큰 가중치를 보이며, 특히 시간축에 대하여 가중치가 다르게 나타남을 보여준다.

[0107] 히트맵에서 가중치가 가장 큰 30 초 ~ 40 초 사이의 구간의 로그 멜-스펙트로그램(log-mel spectrogram)을 확인해보면 묵음의 비율에서 차이가 발생하는 것을 확인할 수 있다.

[0108] 본 예시에서 어텐션 알고리즘을 사용할 경우, 스트레스 판별에 도움이 되는 말의 빠르기 또는 묵음의 비중, 억양의 변화 등이 잘 드러나는 시간적 구간을 찾아 가중치를 부여함으로써, 효과적으로 스트레스 판별을 수행할 수 있다.

[0109] 이상의 설명은 본 발명의 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명의 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명의 실시예들은 본 발명의 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0110] 100: 스트레스 판별 장치

110: 입력부 120: 출력부

130: 프로세서 140: 메모리

150: 데이터 베이스

210: 학습용 음성 신호 획득부 220: 제1 특징벡터 추출부

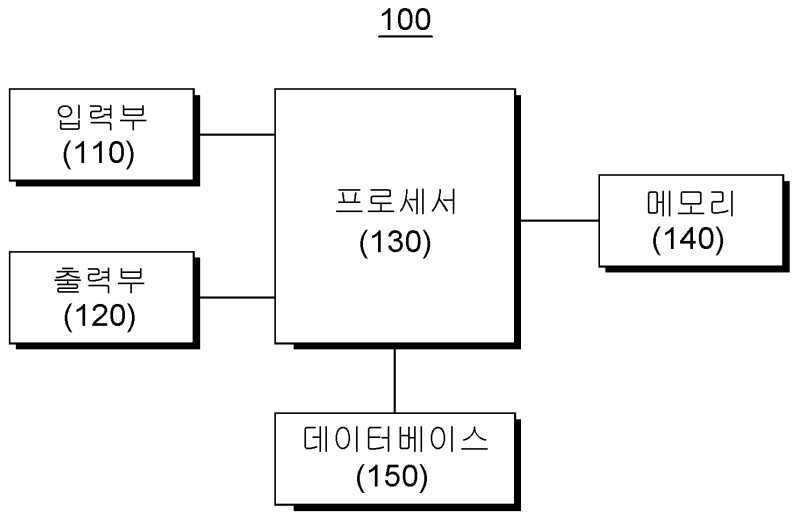
230: 모델 학습부 240: 심층 신경망 모델 처리부

250: 테스트용 음성 신호 획득부 260: 제2 특징벡터 추출부

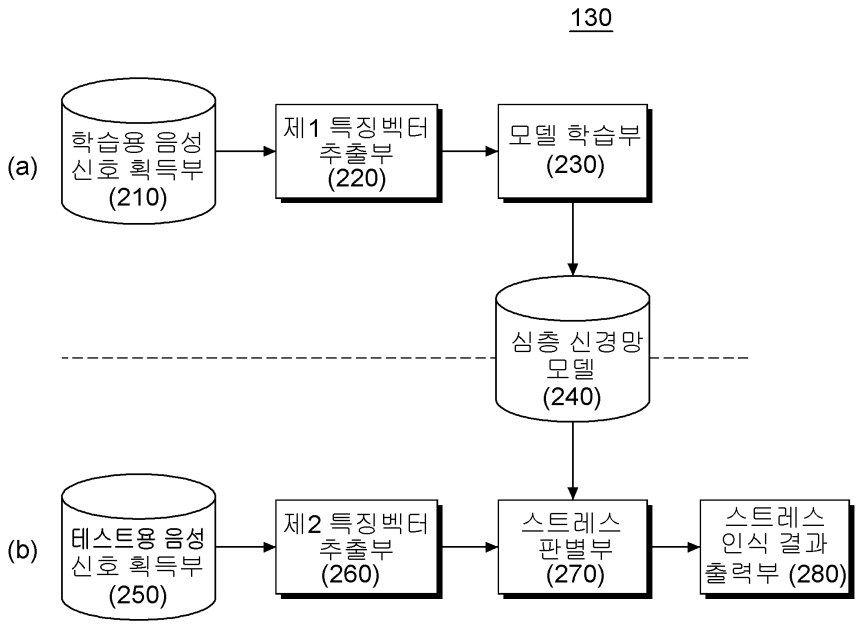
270: 스트레스 판별부 280: 스트레스 인식 결과 출력부

도면

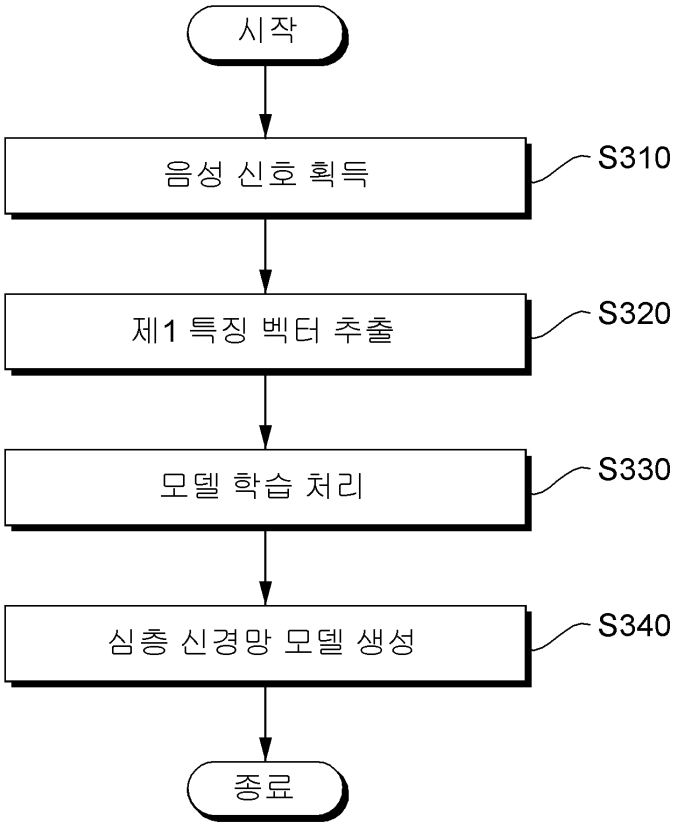
도면1



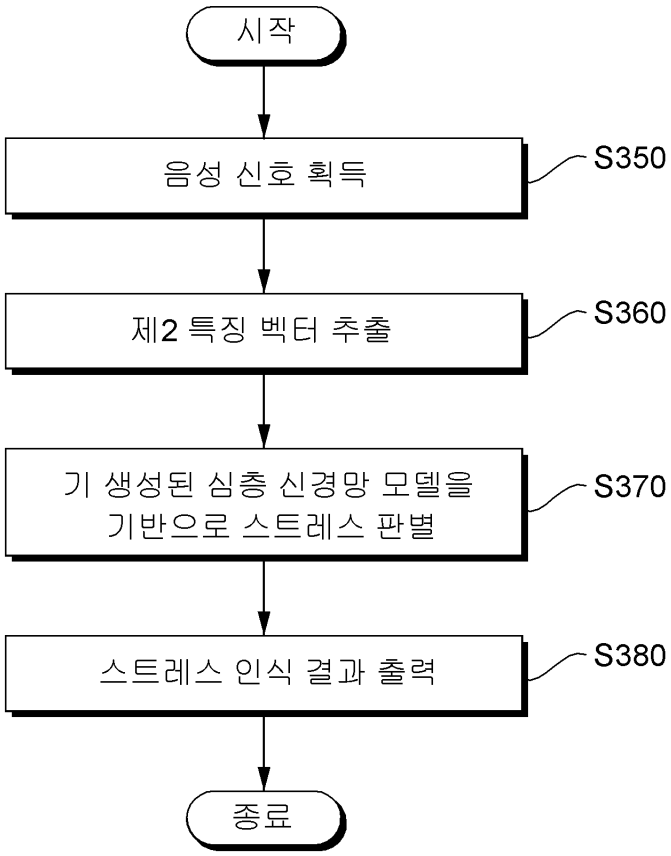
도면2



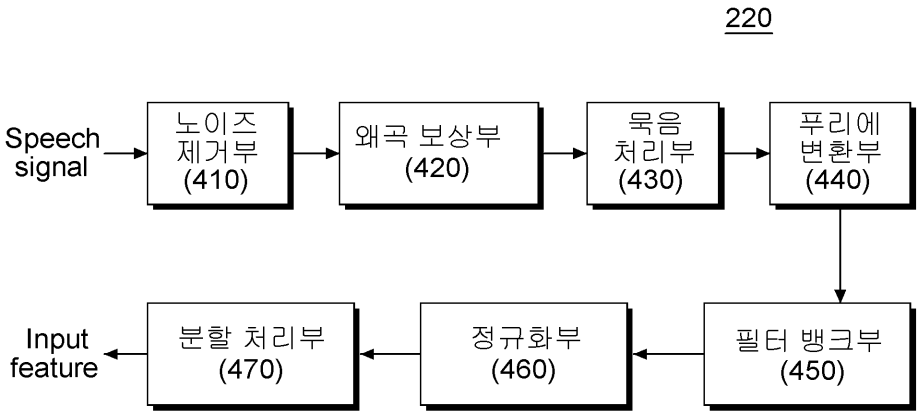
도면3a



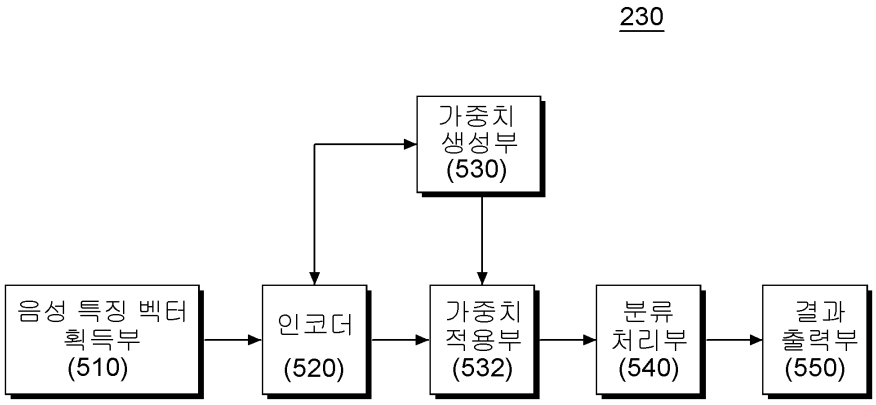
도면3b


시작
음성 신호 획득
S350
제2 특징 벡터 추출
S360
기 생성된 심층 신경망 모델을
기반으로 스트레스 판별
S370
스트레스 인식 결과 출력
S380
종료

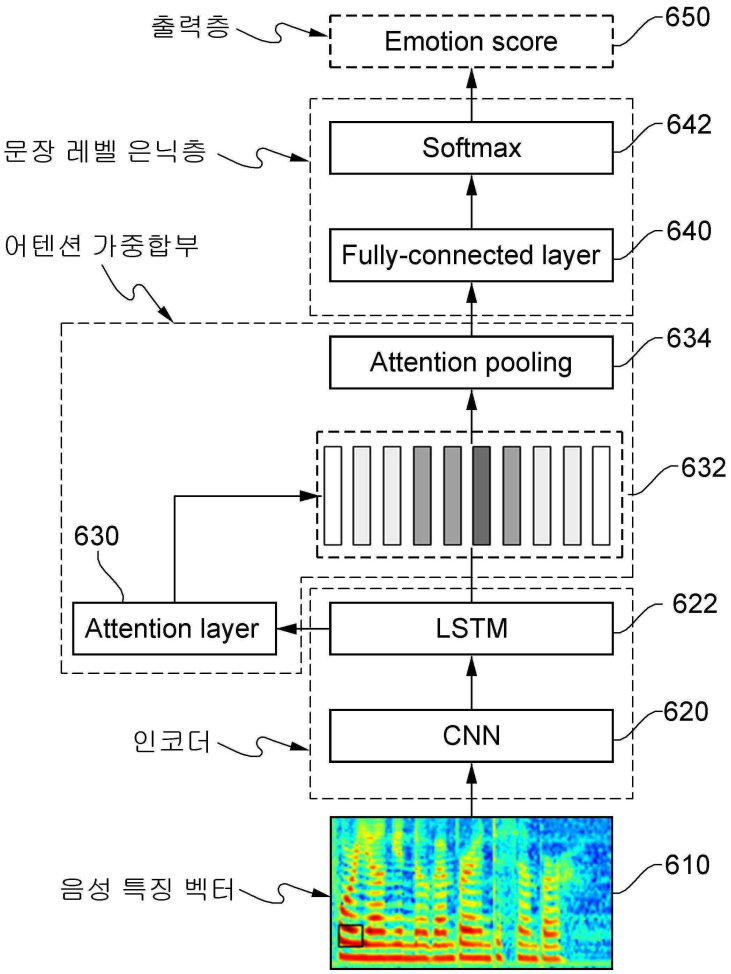
도면4


220
Speech signal
노이즈 제거부 (410)
왜곡 보상부 (420)
묵음 처리부 (430)
푸리에 변환부 (440)
필터 뱅크부 (450)
정규화부 (460)
분할 처리부 (470)
Input feature

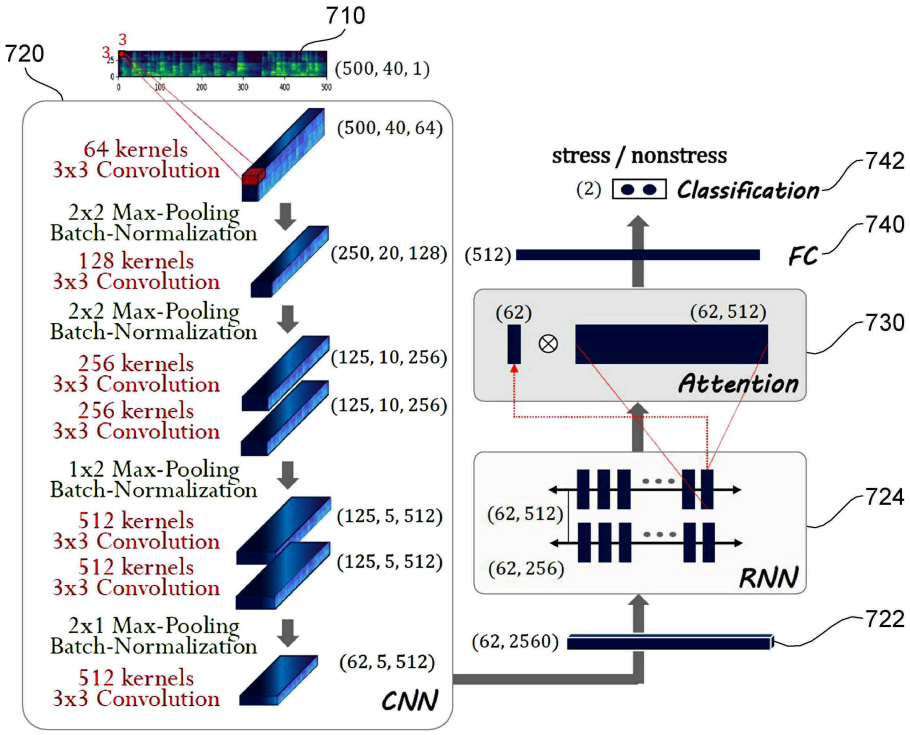
도면5


230
가중치
생성부
(530)
음성 특징 벡터
획득부
(510)
인코더
(520)
가중치
적용부
(532)
분류
처리부
(540)
결과
출력부
(550)

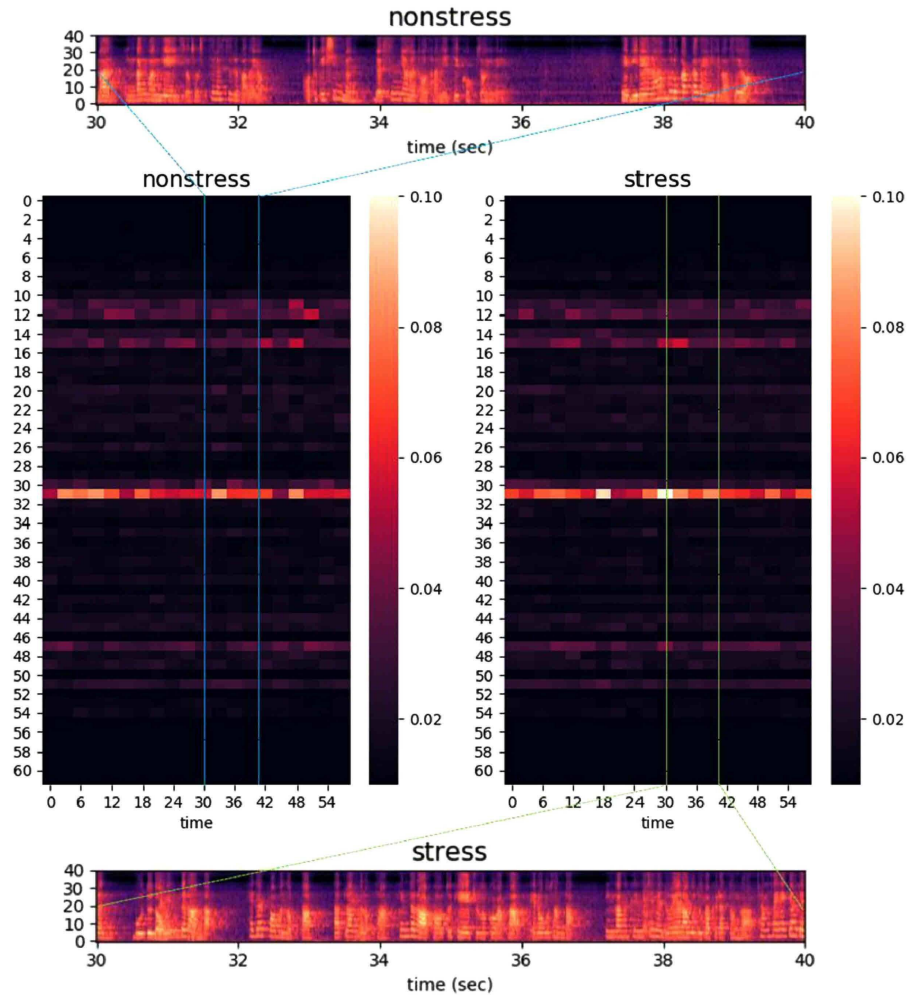
도면6


출력층
Emotion score
650
문장 레벨 은닉층
Softmax
642
어텐션 가중합부
Fully-connected layer
640
Attention pooling
634
632
630
Attention layer
LSTM
622
인코더
CNN
620
음성 특징 벡터
610

도면7


710
720
(500, 40, 1)
64 kernels
3x3 Convolution
(500, 40, 64)
2x2 Max-Pooling
Batch-Normalization
128 kernels
3x3 Convolution
(250, 20, 128)
2x2 Max-Pooling
Batch-Normalization
256 kernels
3x3 Convolution
(125, 10, 256)
256 kernels
3x3 Convolution
(125, 10, 256)
1x2 Max-Pooling
Batch-Normalization
512 kernels
3x3 Convolution
(125, 5, 512)
512 kernels
3x3 Convolution
(125, 5, 512)
2x1 Max-Pooling
Batch-Normalization
512 kernels
3x3 Convolution
(62, 5, 512)
CNN
stress / nonstress
(2)
Classification
742
(512)
FC
740
(62)
(62, 512)
Attention
730
(62, 512)
(62, 256)
RNN
724
(62, 2560)
722

도면8


nonstress
time (sec)
nonstress
time
stress
time
stress
time (sec)