(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년06월29일
(11) 등록번호 10-2549122
(24) 등록일자 2023년06월26일

(51) 국제특허분류(Int. Cl.)
G10L 25/63 (2013.01) *G10L 15/14* (2006.01)
G10L 15/16 (2006.01) *G10L 19/02* (2006.01)
(52) CPC특허분류
G10L 25/63 (2013.01)
G10L 15/14 (2013.01)
(21) 출원번호 10-2021-0088522
(22) 출원일자 2021년07월06일
심사청구일자 2021년07월06일
(65) 공개번호 10-2023-0007781
(43) 공개일자 2023년01월13일
(56) 선행기술조사문헌
Mei,Xiaoguang et.al, Spectral-Spatial Attention Networks for Hyperspectral Image Classification, Remote sensing, MDPI AG, Apr. 2019, Vol.11, no.8, pp.963*
(뒷면에 계속)

(73) 특허권자
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
권순일
서울특별시 강남구 압구정로 201, 77동 1402호 (압구정동, 현대아파트)
무스타킴
서울특별시 광진구 동일로42길 14 (군자동)
(74) 대리인
특허법인엠에이피에스

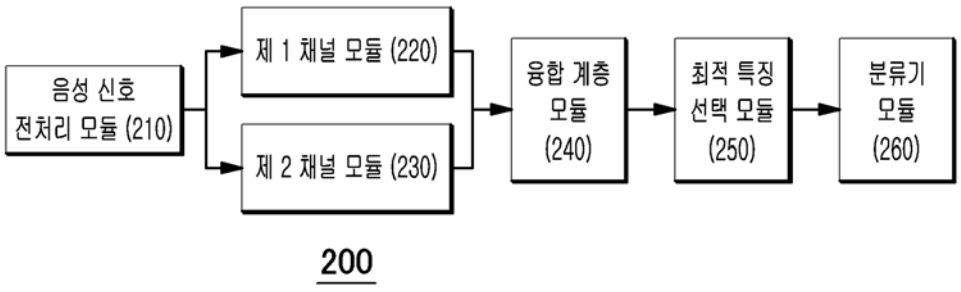
전체 청구항 수 : 총 12 항 심사관 : 김영신

(54) 발명의 명칭 **음성 신호에 기반한 발화자의 감정 인식 장치 및 방법**

(57) 요 약

본 발명의 일 측면에 따른 음성 신호에 기반하여 발화자의 감정을 인식하는 감정 인식 장치는 음성 기반 감정 인식 프로그램이 저장된 메모리; 및 상기 메모리에 저장된 프로그램을 실행하는 프로세서를 포함한다. 음성 기반 감정 인식 프로그램은, 발화자의 음성 신호를 수신하고, 수신한 음성 신호를 감정 분류 모델에 입력하여 발화자의 감정을 분류하되, 상기 감정 분류 모델은 상기 음성 신호의 스펙트럼으로부터 스펙트럼 특징을 추출하는 제 1 채널 모듈, 상기 음성 신호의 스펙트로그램으로부터 공간 특징을 추출하는 제 2 채널 모듈, 상기 제 1 채널 모듈에서 출력된 스펙트럼 특징과 상기 제 2 채널 모듈에서 출력된 공간 특징으로부터, 공동 공간스펙트럼 특징 벡터를 생성하는 융합 계층 모듈, 상기 융합 계층 모듈의 출력으로부터 최적의 특징을 선택하는 최적 특징 선택 모듈 및 상기 선택된 최적의 특징에 대하여 감정 분류를 수행하는 분류기 모듈을 포함한다.

대 표 도 - 도2



(52) CPC특허분류
G10L 15/16 (2013.01)
G10L 19/02 (2013.01)

(56) 선행기술조사문헌
Mustaqeem et.al, Att-Net: Enhanced emotion recognition system using lightweight self-attention module, Applied soft computing, Elsevier, Apr. 2021, Vol.102, pp.107101*
Mustaqeem et.al, Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, IEEE access : practical research, open solutions, IEEE, 2020, Vol.8, pp.79861-79875*
Mustaqeem,Soonil Kwon, Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network, International Journal of Intelligent Systems, May. 2021*
Ozyurt, Fatih , Novel Multi Center and Threshold Ternary Pattern Based Method for Disease Detection Method Using Voice, IEEE access : practical research, open solutions, IEEE, 2020.05, Vol.8, pp.84532-84540*
*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업
과제고유번호 1711118600
과제번호 2020R1F1A1060659
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 음성신호 기반 감정인식 성능향상을 위한 맞춤형 특징요소 그룹 및 딥-네트워크 설계
기 여 율 1/1
과제수행기관명 세종대학교
연구기간 2021.03.01 ~ 2022.02.28
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

음성 신호에 기반하여 발화자의 감정을 인식하는 감정 인식 장치에 있어서,

음성 기반 감정 인식 프로그램이 저장된 메모리; 및

상기 메모리에 저장된 프로그램을 실행하는 프로세서를 포함하며,

상기 음성 기반 감정 인식 프로그램은, 발화자의 음성 신호를 수신하고, 수신한 음성 신호를 감정 분류 모델에 입력하여 발화자의 감정을 분류하되,

상기 감정 분류 모델은 상기 음성 신호의 스펙트럼으로부터 스펙트럼 특징을 추출하는 제 1 채널 모듈, 상기 음성 신호의 스펙트로그램으로부터 공간 특징을 추출하는 제 2 채널 모듈, 상기 제 1 채널 모듈에서 출력된 스펙트럼 특징과 상기 제 2 채널 모듈에서 출력된 공간 특징으로부터, 공동 공간스펙트럼 특징 벡터를 생성하는 융합 계층 모듈, 상기 융합 계층 모듈의 출력으로부터 최적의 특징을 선택하는 최적 특징 선택 모듈 및 상기 선택된 최적의 특징에 대하여 감정 분류를 수행하는 분류기 모듈을 포함하는 것인, 감정 인식 장치.

청구항 2

제 1 항에 있어서,

상기 감정 분류 모델은 상기 음성 신호에 대한 스펙트럼과 스펙트로그램을 생성하는 전처리모듈을 더 포함하는 것인, 감정 인식 장치.

청구항 3

제 1 항에 있어서,

상기 제 1 채널 모듈은 1차원 CNN(convolutional neural network)을 이용하여 상기 스펙트럼 특징을 추출하고,

상기 제 2 채널 모듈은 2차원 CNN을 이용하여 상기 공간 특징을 추출하는 것인, 감정 인식 장치.

청구항 4

제 1 항에 있어서,

상기 융합 계층 모듈은 하기의 수학식에 따라 상기 공동 공간스펙트럼 특징 벡터를 생성하는 것인, 감정 인식 장치.

[수학식 1]

$$F^{(n)} = \int\{w_2 \cdot \int[w_1 \cdot (F_1(R_n) \bowtie F_2(C_n)) + b_1] + b_2\}$$

$F_1(R_n)$은 스펙트럼 특징을 나타내고, $F_2(C_n)$은 공간 특징을 나타내고,W_1과 W_2는 가중치 행렬을 나타내고, b_1과 b_2는 연산중의 연결된 계층에서의 편향들을 나타내고, $\bowtie$ 는 결합 연산자로서, 공간 특징과 스펙트럼 특징을 모두 연결하며, 공동 공간 스펙트럼 특징 벡터는 $F^{(n)}$ 으로 정의됨

청구항 5

제 1 항에 있어서,

상기 최적 특징 선택 모듈은 반복적 이웃 구성 분석 (INCA, Iterative Neighbor Component Analysis) 방법에 따라 최적 특징을 선택하되, 수학식 2에 따라 공동 공간 스펙트럼 특징 벡터를 정규화하고, 수학식 3에 따라 인

덱스를 정렬하는 것인, 감정 인식 장치.

[수학식 2]

$$x(:,i) = \frac{x(:,\ i) - \min(x(:,\ i))}{\max\ (x(:,\ i) - \min\ (x(:,\ i))}$$

i는 1 부터 n 까지의 자연수를 나타내고, 각 특징은 개별적으로 정규화된 후 배열 X에 저장됨

[수학식 3]

$$index = NCA(x, target)$$

인덱스(index)는 정규화된 특징 "x"의 길이로 정렬되고, 대상(target)은 실제 출력을 나타냄

청구항 6

제 1 항에 있어서,

상기 감정 분류 모델은 소프트 맥스 손실함수를 통해 상기 최적 특징 선택 모듈의 출력에 대한 손실을 산출하고, 상기 손실을 최소화하는 방향으로 가중치 업데이트를 수행하는 것인, 감정 인식 장치.

청구항 7

음성 기반 감정 인식 장치를 이용한 감정 인식 방법에 있어서,

발화자의 음성 신호를 수신하는 단계, 및

수신한 음성 신호를 감정 분류 모델에 입력하여 발화자의 감정을 분류하는 단계를 포함하되,

상기 감정 분류 모델은

(a) 상기 음성 신호의 스펙트럼으로부터 스펙트럼 특징을 추출하는 단계;

(b) 상기 음성 신호의 스펙트로그램으로부터 공간 특징을 추출하는 단계;

(c) 상기 스펙트럼 특징과 상기 공간 특징으로부터, 공동 공간 스펙트럼 특징 벡터를 생성하는 단계;

(d) 상기 공동 공간 스펙트럼 특징벡터로부터 최적의 특징을 선택하는 단계; 및

(e) 상기 선택된 최적의 특징에 대하여 감정 분류를 수행하는 단계를 수행하는 것인, 음성 기반 감정 인식 장치를 이용한 감정 인식 방법.

청구항 8

제 7 항에 있어서,

상기 감정 분류 모델은 상기 (a) 단계의 수행전에 상기 음성 신호에 대한 스펙트럼과 스펙트로그램을 생성하는 전처리 단계를 수행하는 것인, 감정 인식 방법.

청구항 9

제 7 항에 있어서,

상기 (a) 단계는 1차원 CNN을 이용하여 상기 스펙트럼 특징을 추출하고,

상기 (b) 단계는 2차원 CNN을 이용하여 상기 공간 특징을 추출하는 것인, 감정 인식 방법.

청구항 10

제 7 항에 있어서,

상기 (c) 단계는 하기의 수학식에 따라 상기 공동 공간스펙트럼 특징 벡터를 생성하는 것인, 감정 인식 방법.

[수학식 1]

$$F^{(n)} = \int\{w_2 \cdot \int[w_1 \cdot (F_1(R_n) \bowtie F_2(C_n)) + b_1] + b_2\}$$

$F_1(R_n)$은 스펙트럼 특징을 나타내고, $F_2(C_n)$은 공간 특징을 나타내고,W_1과 W_2는 가중치 행렬을 나타내고, b_1과 b_2는 연산중의 연결된 계층에서의 편향들을 나타내고, $\bowtie$ 는 결합 연산자로서, 공간 특징과 스펙트럼 특징을 모두 연결하며, 공동 공간 스펙트럼 특징 벡터는 $F^{(n)}$ 으로 정의됨

청구항 11

제 7 항에 있어서,

상기 (d) 단계는 반복적 이웃 구성 분석 (INCA, Iterative Neighbor Component Analysis) 방법에 따라 최적 특징을 선택하되, 수학식 2에 따라 공동 공간 스펙트럼 특징 벡터를 정규화하고, 수학식 3에 따라 인덱스를 정렬하는 것인, 감정 인식 방법.

[수학식 2]

$$x(:,i) = \frac{x(:,\ i) - \min(x(:,\ i))}{\max\ (x(:,\ i) - \min\ (x(:,\ i))}$$

i는 1 부터 n 까지의 자연수를 나타내고, 각 특징은 개별적으로 정규화된 후 배열 X에 저장됨

[수학식 3]

$$index = NCA(x, target)$$

인덱스(index)는 정규화된 특징 "x"의 길이로 정렬되고, 대상(target)은 실제 출력을 나타냄

청구항 12

제 7 항에 있어서,

상기 (e) 단계는 소프트 맥스 손실함수를 통해 최적 특징 선택 모듈의 출력에 대한 손실을 산출하고, 상기 손실을 최소화하는 방향으로 가중치 업데이트를 수행하는 것인, 감정 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 기계 학습 모델을 통해 음성 신호의 스펙트럼 및 스펙트로그램으로부터 발화자의 감정을 인식할 수 있는 음성 기반 감정 인식 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 사람의 음성을 통해 발화자의 감정을 인식하는 기술에 대한 연구가 진행되고 있다. 특히, 인공 지능 기술이나 기계 학습 기술을 이용하는, 스마트 음성 감정 인식(SER, Speech Emotion Recognition) 기술은 디지털 오디오 신호 처리의 새로운 분야로 알려지고 있으며, 인간-컴퓨터 상호 작용(HCI, Human Computer Interface) 기술과 관련된 많은 응용 프로그램에서 중요한 역할을 할것으로 기대하고 있다.

[0003] 기존의 연구는 음성 신호에서 감정인식을 모델링 하기위해 다양한 수의 심층 신경망(DNN)을 도입하고 있다. 예를 들어, 원본 오디오 샘플에서 중요한 신호를 감지하는 DNN 모델이 제안되거나, 오디오 녹음의 특정 표현을 사용하여 모델에 대한 입력을 제공하는 기술이 제안되었다.

[0004] 특히, 기존의 많은 연구들은 효율적인 음성을 이용한 감정인식 시스템을 위해 음성 스펙트로그램, 원 음성 신호

및 log-Mel 스펙트로그램과 같은 다양한 입력 유형의 음성신호를 사용하여 감정인식을 시도해 왔다. 그러나 이러한 방법들은 감정인식을 위해 일부 정보만을 사용하는 모델이며, 완전하지 못한 정보를 시스템에 제공한다. 정확한 음성기반 감정인식 시스템을 만들기 위해서는 다양한 음성관련 정보를 통합적으로 활용하는 것이 중요하며, 본 발명에서는 이와 같이, 다양한 음성관련 정보를 활용하는 감정인식 시스템을 제공하고자 한다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 등록특허공보 제10-1564176호 (발명의 명칭: 감정 인식 시스템 및 그 제어 방법)

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 음성신호의 음성 스펙트럼과 스펙트로그램으로부터 추출되는 특징 정보를 활용하여 음성으로부터 발화자의 감정을 분류할 수 있는 음성 기반 감정 인식 장치 및 방법을 제공하는데 목적이 있다.

[0007] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0008] 상술한 기술적 과제를 해결하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 음성 신호에 기반하여 발화자의 감정을 인식하는 감정 인식 장치는 음성 기반 감정 인식 프로그램이 저장된 메모리; 및 상기 메모리에 저장된 프로그램을 실행하는 프로세서를 포함한다. 음성 기반 감정 인식 프로그램은, 발화자의 음성 신호를 수신하고, 수신한 음성 신호를 감정 분류 모델에 입력하여 발화자의 감정을 분류하되, 상기 감정 분류 모델은 상기 음성 신호의 스펙트럼으로부터 스펙트럼 특징을 추출하는 제 1 채널 모듈, 상기 음성 신호의 스펙트로그램으로부터 공간 특징을 추출하는 제 2 채널 모듈, 상기 제 1 채널 모듈에서 출력된 스펙트럼 특징과 상기 제 2 채널 모듈에서 출력된 공간 특징으로부터, 공동 공간스펙트럼 특징 벡터를 생성하는 융합 계층 모듈, 상기 융합 계층 모듈의 출력으로부터 최적의 특징을 선택하는 최적 특징 선택 모듈 및 상기 선택된 최적의 특징에 대하여 감정 분류를 수행하는 분류기 모듈을 포함하는 것이다.

[0009] 또한, 본 발명의 다른 측면에 따른 음성 기반 감정 인식 장치를 이용한 음성 기반 감정 인식 장치를 이용한 감정 인식 방법은 발화자의 음성 신호를 수신하는 단계 및 수신한 음성 신호를 감정 분류 모델에 입력하여 발화자의 감정을 분류하는 단계를 포함한다. 이때, 감정 분류 모델은 음성 신호의 스펙트럼으로부터 스펙트럼 특징을 추출하는 단계; 음성 신호의 스펙트로그램으로부터 공간 특징을 추출하는 단계; 스펙트럼 특징과 상기 공간 특징으로부터, 공동 공간 스펙트럼 특징 벡터를 생성하는 단계; 공동 공간 스펙트럼 특징벡터로부터 최적의 특징을 선택하는 단계; 및 선택된 최적의 특징에 대하여 감정 분류를 수행하는 단계를 수행한다.

발명의 효과

[0010] 전술한 본원의 과제 해결 수단에 의하면, 음성 신호에 포함된 스펙트럼 특징과 공간적 특징을 효과적으로 추출하여, 발화자의 감정을 자동으로 분류할 수 있다.

[0011] 종래의 경우, 스펙트럼 또는 스펙트로그램에서 추출되는 단일 특징만으로 감정을 분류하기 때문에, 언어 정보 손실로 인한 어려움을 겪었다. 본 발명에서는 스펙트럼 분석 및 공간 특징 분석을 모두 사용하기 때문에, 음성 신호의 정보를 최대한 활용하여 감정 인식을 수행할 수 있다.

[0012] 또한, 음성 신호에서 추출되는 스펙트럼 특징과 공간 특징에 기초하여 공동 공간 스펙트럼 특징 벡터를 생성하고, 그로부터 최적의 특징을 선택하기 위한 알고리즘을 수행하기 때문에, 중복성을 제거하고 최적의 특징을 선택할 수 있다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일 실시예에 따른 음성 기반 감정 인식 장치의 구성을 도시한 블록도이다.

도 2는 본 발명의 일 실시예에 따른 음성 기반 감정 인식 프로그램에 포함된 감정 분류 모델의 구성을 도시한 블록도이다.

도 3은 본 발명의 일 실시예에 따른 감정 분류 모델의 구체적인 구성을 도시한 순서도이다.

도 4는 본 발명의 일 실시예에 따른 감정 인식 방법을 도시한 순서도이다.

도 5는 본 발명의 일 실시예에 따른 감정 분류 모델의 성능 평가 결과를 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

[0014] 아래에서는 첨부한 도면을 참조하여 본원이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본원의 실시예를 상세히 설명한다. 그러나 본원은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0015] 본원 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다.

[0016] 본원 명세서 전체에서, 어떤 부재가 다른 부재 “상에” 위치하고 있다고 할 때, 이는 어떤 부재가 다른 부재에 접해 있는 경우뿐 아니라 두 부재 사이에 또 다른 부재가 존재하는 경우도 포함한다.

[0017] 이하 첨부된 도면을 참고하여 본 발명의 일 실시예를 상세히 설명하기로 한다.

[0018] 도 1은 본 발명의 일 실시예에 따른 음성 기반 감정 인식 장치의 구성을 도시한 블록도이고, 도 2는 본 발명의 일 실시예에 따른 음성 기반 감정 인식 프로그램에 포함된 감정 분류 모델의 구성을 도시한 블록도이다.

[0019] 도시된 바와 같이 음성 기반 감정 인식 장치(100)는 통신 모듈(110), 메모리(120), 프로세서(130) 및 데이터베이스(140)를 포함할 수 있다. 또한, 음성 기반 감정 인식 장치(100)는 마이크 등을 내장할 수 있으며, 이를 통해 직접 음성 신호를 생성하는 것도 가능하다.

[0020] 음성 기반 감정 인식 장치(100)는 각 사용자 단말로부터 음성 신호를 수신하고, 그로부터 감정 분류 결과를 제공하는 서버로서 동작할 수 있다. 이러한 경우, 음성 기반 감정 인식 장치(100)는 SaaS (Software as a Service), PaaS (Platform as a Service) 또는 IaaS (Infrastructure as a Service)와 같은 클라우드 컴퓨팅 서비스 모델에서 동작할 수 있다. 또한, 음성 기반 감정 인식 장치(100)는 사설(private) 클라우드, 공용(public) 클라우드 또는 하이브리드(hybrid) 클라우드와 같은 형태로 구축될 수 있다.

[0021] 통신 모듈(110)은 발화자의 음성 신호를 외부 기기로부터 수신하는 것으로서, 각종 스마트 단말에 연결된 마이크 등을 통해 입력된 음성 신호를 통신망(300)을 통해 수신할 수 있다. 또한 통신 모듈(110)은 각종 외부 장치(서버 또는 단말) 로부터 음성 기반 감정 인식 프로그램 등의 업데이트 정보 등을 수신하여 프로세서(130)로 전송할 수 있다.

[0022] 통신 모듈(110)은 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.

[0023] 메모리(120)에는 발화자의 음성을 기반으로 발화자의 감정을 분류하는 음성 기반 감정 인식 프로그램이 저장된다. 이러한 메모리(120)에는 음성 기반 감정 인식 장치(100)의 구동을 위한 운영 체제나 음성 기반 감정 인식 프로그램의 실행 과정에서 발생되는 여러 종류가 데이터가 저장된다.

[0024] 이때, 음성 기반 감정 인식 프로그램은, 발화자의 음성 신호를 통신 모듈(110)을 통해 수신하고, 수신한 음성 신호를 감정 분류 모델(200)에 입력하여 발화자의 감정을 분류한다. 이때, 감정 분류 모델(200)은 음성 신호로부터 스펙트럼 특징을 추출하는 제 1 채널 모듈(220), 음성 신호의 스펙트로그램으로부터 공간 특징을 추출하는 제 2 채널 모듈(230), 제 1 채널 모듈(220)에서 출력된 스펙트럼 특징과 제 2 채널 모듈(230)에서 출력된 공간 특징으로부터, 공동 공간스펙트럼 특징 벡터를 생성하는 융합 계층 모듈(240), 융합 계층 모듈(240)의 출력으로부터 최적의 특징을 선택하는 최적 특징 선택 모듈(250) 및 선택된 최적의 특징에 대하여 감정 분류를 수행하는 분류기 모듈(260)을 포함한다. 음성 기반 감정 인식 프로그램의 구체적인 내용에 대해서는 추후 설명하기로 한다.

[0025] 이때, 메모리(120)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치를 통칭하는 것이다.

[0026] 또한, 메모리(120)는 프로세서(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행할 수 있다. 여기서, 메모리(120)는 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치 외에 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0027] 프로세서(130)는 메모리(120)에 저장된 프로그램을 실행하되, 음성 기반 감정 인식 프로그램의 실행에 따라, 감정 분류 모델의 구축 과정과 구축된 감정 분류 모델을 통해 음성을 기반으로 발화자의 감정을 분류하는 작업을 수행한다.

[0028] 이러한 프로세서(130)는 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 예를 들어 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 이와 같이 하드웨어에 내장된 데이터 처리 장치의 일 예로써, 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0029] 데이터베이스(140)는 프로세서(130)의 제어에 따라, 음성 기반 감정 인식 장치(100)에 필요한 데이터를 저장 또는 제공한다. 이러한 데이터베이스(140)는 메모리(120)와는 별도의 구성 요소로서 포함되거나, 또는 메모리(120)의 일부 영역에 구축될 수도 있다.

[0030] 한편, 음성 기반 감정 인식 장치(100)는 장치(100)에 내장되거나 이에 접속된 마이크 등을 통해 발화자의 음성 신호를 녹음하여, 음성 신호를 직접 생성하는 동작도 수행할 수 있으며, 이에 대해 감정 인식을 수행할 수 있다.

[0031] 도 3은 본 발명의 일 실시예에 따른 감정 분류 모델의 구체적인 구성을 도시한 순서도이다.

[0032] 도 2와 도 3을 함께 참조하여 설명하면, 음성 신호가 수신되면, 음성 신호 전처리 모듈(210)를 통해 음성 신호에 대한 스펙트럼과 스펙트로그램을 각각 생성하는 전처리를 수행한다.

[0033] 먼저, 음성 신호를 입력으로 하고, 아래 수학식 1에 정의된 RMS(Root Mean Square) 함수를 통해 정규화를 수행한다.

[0034] [수학식 1]

[0035] $$R = \sqrt{\frac{1}{n}[(fs_1)^2 + (fs_1)^2 + \cdots (fs_n)^2]}$$

[0036] RMS의 전체 결과는 "R"로 표시되고, 음성 신호의 스케일링 인자는 "f"로 표시되며, 음성 신호의 진폭 변화는 "s"로 직접 수행된다.

[0037] RMS 함수를 통해 정규화된 음성 세그먼트는 스펙트럼 분석을 위해 제 1 채널 모듈(220)로 전송된다.

[0038] 또한, 전처리 모듈(210)은 단시간 푸리에 변환(STFT)를 이용하여, 정규화된 음성 세그먼트를 스펙트로그램으로 변환하고, 이를 제 2 채널 모듈(230)로 전송한다. 스펙트로그램은 음성 신호를 2차원 이미지 형태로 시각적으로 표현한 것으로, 2D CNN 모델이 다양한 감정을 인식하기 위해 높은 수준의 특징을 추출하는 데 가장 적합하다.

[0039] 제 1 채널 모듈(220)은 전처리 모듈(210)로부터 수신한 스펙트럼 음성 신호로부터 스펙트럼 특징을 추출한다. 본 발명에서는 제 1 채널 모듈(220)로서 CNN(convolutional neural network)을 사용한다.

[0040] 제 1 채널 모듈(220)은 확장된 CNN(dilated CNN)을 사용할 수 있다. 제 1 채널 모듈(220)은 음성 신호의 n 번째 세그먼트(Rn)에 대한 스펙트럼을 입력으로서 수신하고, 해당 세그먼트에 대한 스펙트럼 특징(F1(Rn))을 추출하는데, 이를 위해 아래의 수학식 2를 사용할 수 있다.

[0041] [수학식 2]

[0042] $$z_i^L = \sum_j \int (w_{i,j}^L * z_j^{L-1} + b_i^L)$$

[0043] 이때, z_j^{L-1} 은 z_i^L 특징맵에 대하여 카테고리적으로 연결된 네트워크 레이어(L-1)에서의 특징맵을 나타내고, $w_{i,j}^L$ 은 z_j^{L-1} 에 대한 컨볼루션 커널을 나타내고, b_i^L 는 바이어스, *는 컨볼루션 연산자, $\int(.)$는 Relu나 시그모이드와 같은 비선형 활성함수를 나타낸다.

[0044] 제 1 채널 모듈(220)에 입력되는 입력 텐서(Rn)는 1차원 신호이므로, 1차원 컨볼루션 및 풀링 레이어를 사용하여, 1차원 계산을 수행하고, 그에 대한 스펙트럼 특징을 추출한다.

[0045] 제 2 채널 모듈(230)은 전처리 모듈(210)로부터 수신한 음성 신호의 스펙트로그램으로부터 공간 특징을 추출한다. 본 발명에서는 제 2 채널 모듈(230)로서 2차원 CNN을 사용한다.

[0046] 제 2 채널 모듈(230)은 확장된 CNN(dilated CNN)을 사용할 수 있다. 제 2 채널 모듈(230)은 입력 스펙트로그램(Cn)을 수신하고, 공간 특징(F2(Cn))을 추출한다. 이때, 제 2 채널 모듈(230)은 도시된 바와 같이, 확장된 CNN 계층 및 BN(Batch Normalization) 계층 서로 교호하면서 복수개가 배치된 구성과, 폴링 계층을 포함하여 이루어질 수 있다.

[0047] 융합 계층 모듈(240)은 제 1 채널 모듈(220)에서 출력된 스펙트럼 특징과 제 2 채널 모듈(230)에서 출력된 공간 특징으로부터, 공동 공간 스펙트럼 특징(Spatial & sepctral fused features)을 나타내는 공동 공간 스펙트럼 특징 벡터를 생성한다. 이를 위해 아래의 수학식 3을 사용할 수 있다.

[0048] [수학식 3]

[0049] $$F^{(n)} = \int\{w_2 \cdot \int[w_1 \cdot (F_1(R_n) \bowtie F_2(C_n)) + b_1] + b_2\}$$

[0050] W_1과 W_2는 가중치 행렬을 나타내고, b_1과 b_2는 연산중의 연결된 계층에서의 편향들을 나타낸다. $\bowtie$ 는 결합 연산자로서, 공간 특징과 스펙트럼 특징을 모두 연결하며, 공동 공간 스펙트럼 특징 벡터는 $F^{(n)}$으로 정의된다.

[0051] 최적 특징 선택 모듈(250)은 융합 계층 모듈의 출력으로부터 최적의 특징을 선택한다. 최적 특징 선택 모듈(250)은 여러 가지 특징 선택 알고리즘에 의해 구현될 수 있다.

[0052] 예를 들어, NCA (Neighbor Component Analysis) 방법이 알려져 있다. NCA는 다변량 데이터를 여러 클래스로 분류하는 지도 학습 방법으로서, 거리 기반 방법이며 양의 가중치 특징을 선택하는 방법으로서 많이 사용되고 있는데, 중복성과 불일치로 인해 분류기가 최적의 특징을 선택하기 어렵게하는 문제가 있다. 이러한 문제를 해결하기 위해 중복성을 자동으로 제거하고 최적의 특징 개수를 선택하는 반복적 이웃 구성 분석 (INCA, Iterative Neighbor Component Analysis)을 제안하였다.

[0053] INCA는 작업(task)에 따라 특징 벡터의 길이를 선택하는데, 이는 NCA 특징 선택 방법을 활용하는 것으로서, 다음과 같은 단계를 수행한다.

[0054] 먼저, 수학식 4에 의해 NCA를 효과적으로 사용하는 최소-최대 정규화 방법을 적용하여 공동공간 스펙트럼 특징 벡터를 정규화한다.

[0055] [수학식 4]

[0056] $$x(:,i) = \frac{x(:,i) - \min(x(:,i))}{\max(x(:,i)) - \min(x(:,i))}$$

[0057] 이때, i는 1 부터 n 까지의 자연수를 나타내고, 각 특징은 개별적으로 정규화된 후 배열 X에 저장된다. 정규화 이후 NCA에 따라 수학식 5를 사용하여 인덱스를 정렬한다.

[0058] [수학식 5]

[0059] $$index = NCA(x, target)$$

[0060] 이 방정식을 사용하여 인덱스는 정규화된 특징 "x"의 길이로 정렬되고 대상(target)은 실제 출력을 나타낸다. 인덱스를 사용한 반복적 특징 선택은 정의되지 않은 특징 범위로 인해 계산 복잡성이 높다. 비용을 줄이기 위해 "x = {1, 2, 3… 1000}"과 같이 특징 개수를 제한한 다음 선택한 특징들에서 최소 오류율을 찾는다. 마지막으로 인덱스 값을 사용하여 최적의 특징을 선택하고 최종 예측을 위한 추가 프로세스를 수행한다.

[0061] 분류기 모듈(260)은 선택된 최적의 특징에 대하여 감정 분류를 수행한다. 예를 들면, 최적의 특징에 대한 출력으로서 발화자의 감정 상태를 '화남', '슬픔', '행복', '보통' 등으로 분류할 수 있다. 이를 위해, 분류기 모듈(260)은 복수의 학습 데이터를 이용하여 반복적으로 학습되며, 다음과 같은 수학식 6에 의해 정의되는 소프트맥스 손실 함수를 사용하여, 최적 특징 선택 모듈(250)의 출력에 대한 손실을 산출하고, 손실을 최소화하는 방향으로 가중치 업데이트를 수행한다.

[0062] [수학식 6]

[0063] $$J(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{L} 1\{k = L^{(n)}\} \log C_k^{(n)}$$

[0064] N 는 훈련 데이터의 전체 개수를 나타내고, $L^{(n)}$은 n번째 훈련 데이터의 실측 레이블(ground truth label)을 나타내고, $C_k^{(n)}$ 은 $C^{(n)}$의 k 번째 요소를 나타내는 것으로, n번째 훈련 데이터에 대한 k 번째 감정의 확률을 나타낸다. Θ 는 커널 및 편향 값을 나타낸다. 1 은 노출 함수(revealing function)로서 괄호 안의 조건이 충족되면 값이 1이되고 그렇지 않으면 0이되는 조건을 갖는다.

[0065] 이와 같은 손실함수를 통해, Θ 에 대한 확률적 경사 하강법을 이용하여 최종 예측을 위한 소프트 맥스 계층을 최적화할 수 있다.

[0066] 참고로, 본 발명의 일 실시예에 따라면, 모델 학습 중에 가우스 랜덤 분포를 사용하여 표준 분산 0.05와 평균을 사용하여 컨벌루션 연산의 모든 커널을 초기화했다. 그리고, 100개의 학습 에포크가 있는 모든 데이터 세트에 대해 0.0001의 고정 학습률을 사용했다. 전체 훈련 프로세스에 대해 64 개의 배치 사이즈(batch size)를 선택하고 훈련에서 0.215 손실, 검증에서 0.346 손실로 높은 정확도를 달성했다.

[0067] 높은 인식 결과를 확보하려면 딥 러닝 모델에 충분한 훈련 데이터가 필요하다. 그러나 SER 분야에서는 실제 라벨링된 데이터가 제한되어 모델 학습에 충분하지 않다. 제한된 학습 데이터로 모델 성능을 개선하기 위해 각 발화를 여러 세그먼트로 나누고 학습 중에 동일한 발화의 모든 세그먼트가 포함된 유사한 실제 레이블을 제공하여 학습을 수행하였다.

[0068] 한편, 본 발명에 따른 감정 분류 모델에 대한 평가를 위해, 3가지 표준 데이터 베이스를 사용하였는데, 이는 각각 EMO-DB, SAVEE 및 RAVDESS로 알려진 것들이다.

[0069] 이 모든 데이터베이스는 스크립트로 작성되어 있으며 배우는 두려움, 분노, 슬픔, 놀라움, 행복, 침착함과 같은 다양한 감정으로 준비된 문장을 읽는 방법에 의해 데이터가 수집된 것이다. 각각의 데이터베이스는 복수의 남성 또는 여성 배우들이 녹음한 음성 파일을 포함하며, 이는 소정의 샘플링 율에 따라 녹음된 것들이다.

[0070] 본 발명에서는 이러한 표준 데이터베이스를 활용하여, 감정 분류 모델을 학습하고, 모델을 평가하였다. 보다 구체적으로 살펴보면, 각 데이터를 각 폴드에서 80 : 20 의 비율로 분할하고, 10 폴드 교차 검증 기법을 활용하여 시스템 성능을 평가했다.

[0071] 도 4는 본 발명의 일 실시예에 따른 감정 인식 방법을 도시한 순서도이다.

[0072] 먼저, 감정 인식 장치(100)가 발화자의 음성 신호를 수신한다(S410).

[0073] 다음으로, 수신한 음성 신호를 감정 분류 모델에 입력하여 발화자의 감정을 분류하는데, 다음과 같은 단계를 순차적으로 수행한다.

[0074] 음성 신호에 대한 스펙트럼과 스펙트로그램을 생성하는 전처리 단계를 수행한다. 앞서 살펴본 바와 같이, 스펙

트럼의 생성을 위해, RMS함수를 통해 정규화하는 과정이나, 스펙트로그램의 생성을 위해 단시간 푸리에 변환을 사용할 수 있다.

[0075] 다음으로, 음성 신호의 스펙트럼으로부터 스펙트럼 특징을 추출한다(S420). 이를 위해, 제 1 채널 모듈(220)을 통해 1차원 CNN을 이용하여 스펙트럼 특징을 추출한다.

[0076] 다음으로, 음성 신호의 스펙트로그램으로부터 공간 특징을 추출한다(S430). 이를 위해, 제 2 채널 모듈(230)을 통해 2차원 CNN을 이용하여 공간 특징을 추출한다.

[0077] 다음으로, 스펙트럼 특징과 공간 특징으로부터, 공동 공간 스펙트럼 특징 벡터를 생성한다(S440). 이를 위해, 앞서 설명한 수학식 3을 이용하여 공동 공간 스펙트럼 특징 벡터($\mathbf{F}^{(n)}$)를 생성한다.

[0078] 다음으로, 공동 공간 스펙트럼 특징벡터로부터 최적의 특징을 선택한다(S450). 이를 위해, 앞서 설명한 반복적 이웃 구성 분석 (INCA, Iterative Neighbor Component Analysis) 방법에 따라 최적 특징을 선택한다.

[0079] 다음으로, 선택된 최적의 특징에 대하여 감정 분류를 수행한다(S460). 이를 위해, 소프트 맥스 손실함수를 통해 최적 특징 선택 모듈의 출력에 대한 손실을 산출하고, 손실을 최소화하는 방향으로 가중치 업데이트를 수행하는 동작을 반복수행한다. 그리고, 이와 같이, 가중치 업데이트가 완료된 감정 분류 모델에 대하여, 분류하고자 하는 음성 신호를 입력하여, 감정을 분류하는 추론 과정을 수행한다.

[0080] 도 5는 본 발명의 일 실시예에 따른 감정 분류 모델의 성능 평가 결과를 도시한 것이다.

[0081] 제안된 모델에 대한 전체적인 예측 결과를 그래프로 도시한 결과 EMO-DB, SAVEE, RAVDESS 데이터 셋에 대해 각각 95 %, 82 %, 85 %의 인식률을 확보 할 수 있음을 확인할 수 있다.

[0082] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독 가능 매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0083] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0084] 전술한 본원의 설명은 예시를 위한 것이며, 본원이 속하는 기술분야의 통상의 지식을 가진 자는 본원의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0085] 본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0086] 100: 음성 기반 감정 인식 장치

110: 통신 모듈

120: 메모리

130: 프로세서

140: 데이터베이스

200: 감정 분류 모델

210: 음성 신호 전처리 모듈

220: 제 1 채널 모듈

230: 제 2 채널 모듈

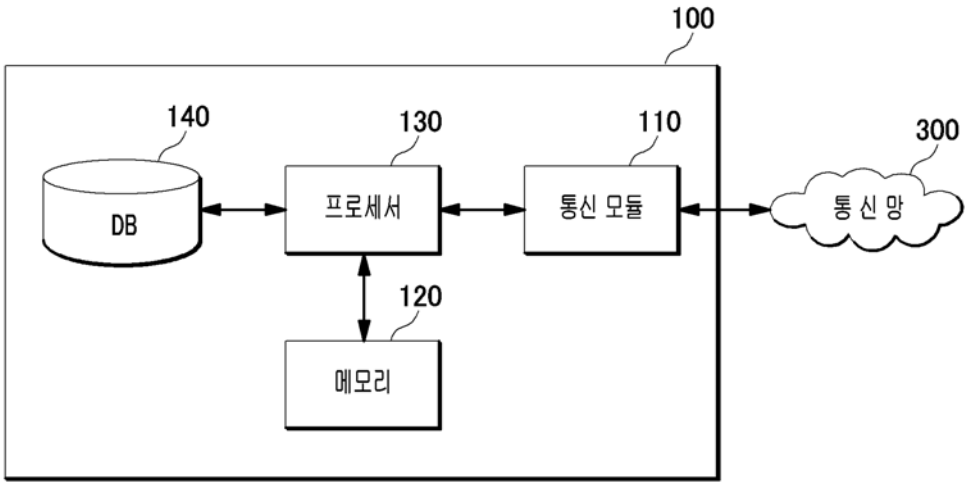
240: 융합 계층 모듈

250: 최적 특징 선택 모듈

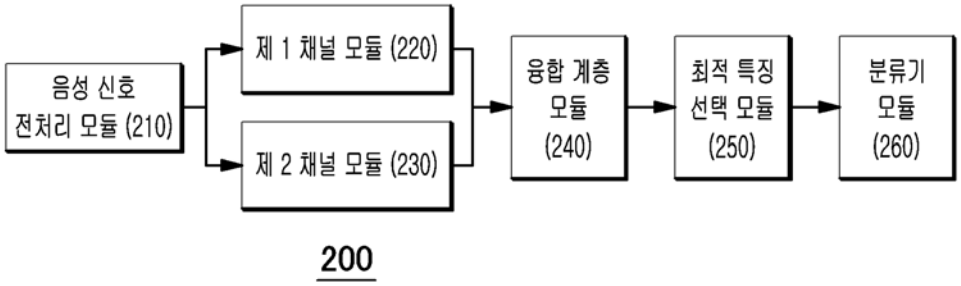
260: 분류기 모듈

도면

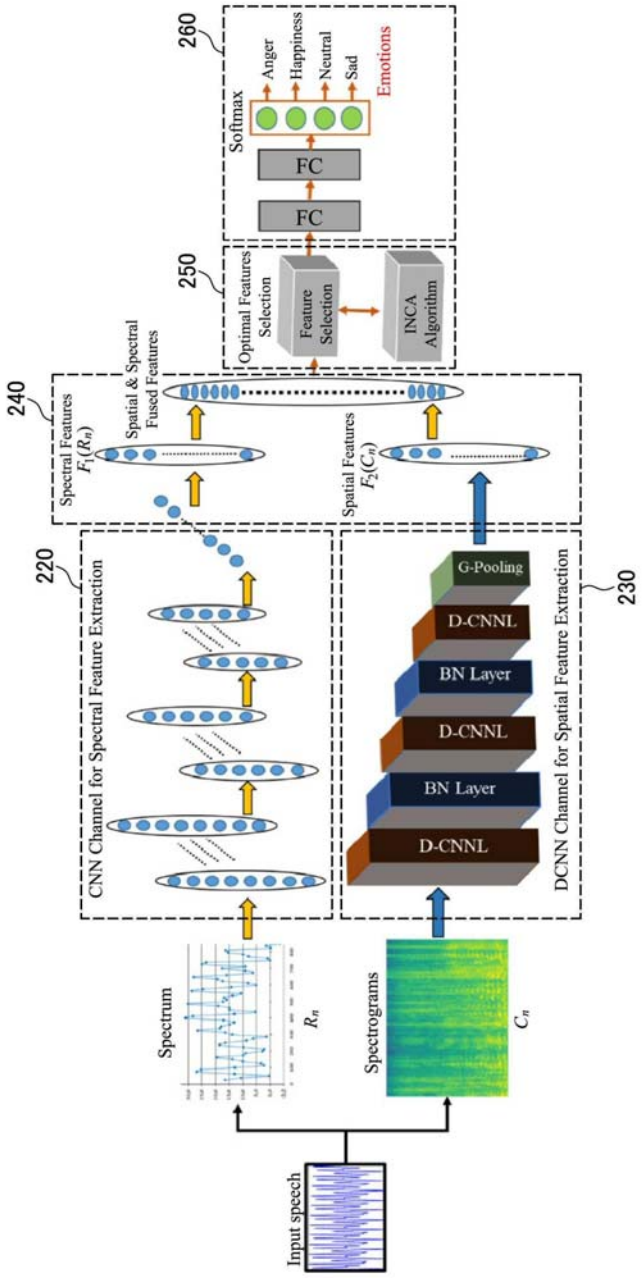
도면1



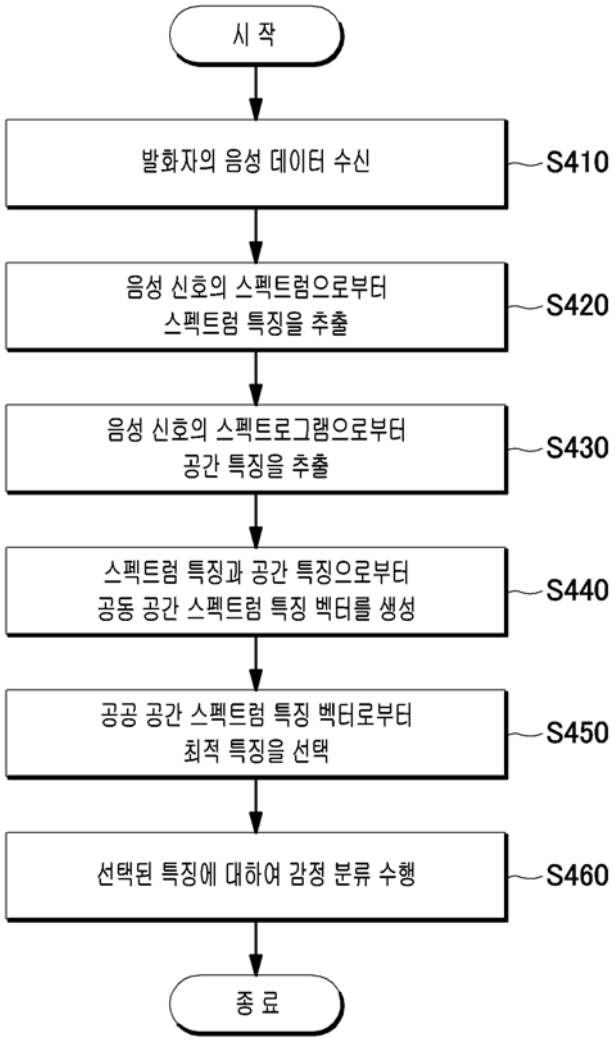
도면2



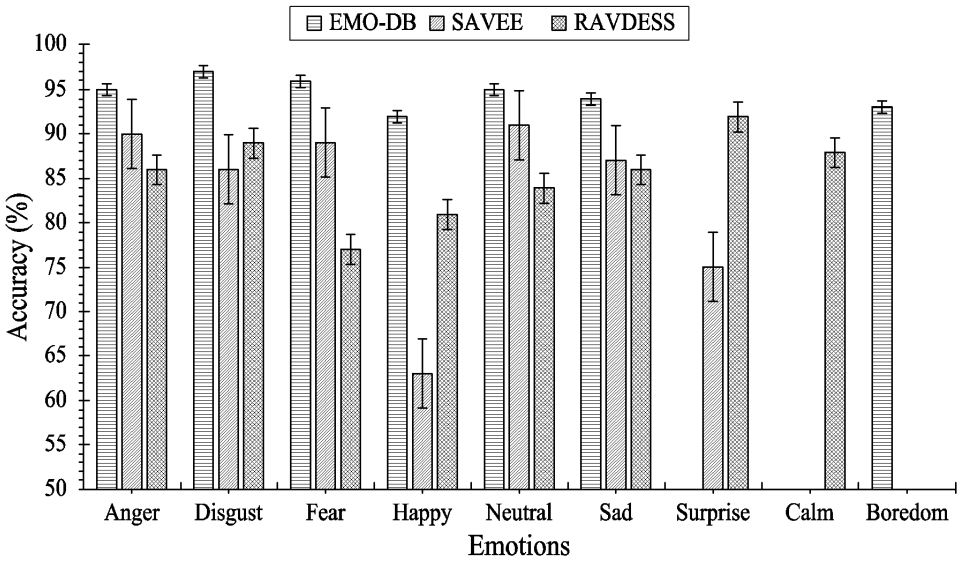
도면3


260
Softmax
Anger
Happiness
Neutral
Sad
Emotions
FC
FC
250
Optimal Features Selection
Feature Selection
INCA Algorithm
240
Spectral Features $F_1(R_n)$
Spatial & Spectral Fused Features
Spatial Features $F_2(C_n)$
220
CNN Channel for Spectral Feature Extraction
230
DCNN Channel for Spatial Feature Extraction
G-Pooling
D-CNNL
BN Layer
D-CNNL
BN Layer
D-CNNL
Spectrum
R_n
Spectrograms
C_n
Input speech

도면4


시 작
발화자의 음성 데이터 수신
S410
음성 신호의 스펙트럼으로부터
스펙트럼 특징을 추출
S420
음성 신호의 스펙트로그램으로부터
공간 특징을 추출
S430
스펙트럼 특징과 공간 특징으로부터
공동 공간 스펙트럼 특징 벡터를 생성
S440
공공 공간 스펙트럼 특징 벡터로부터
최적 특징을 선택
S450
선택된 특징에 대하여 감정 분류 수행
S460
종 료

도면5


EMO-DB
SAVEE
RAVDESS
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
Anger
Disgust
Fear
Happy
Neutral
Sad
Surprise
Calm
Boredom
Emotions