(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년06월12일
(11) 등록번호 10-2541660
(24) 등록일자 2023년06월05일

(51) 국제특허분류(Int. Cl.)
G10L 25/63 (2013.01) *G06F 18/00* (2023.01)
G10L 17/02 (2013.01) *G10L 17/04* (2013.01)
(52) CPC특허분류
G10L 25/63 (2013.01)
G06F 18/253 (2023.01)
(21) 출원번호 10-2021-0000952
(22) 출원일자 2021년01월05일
심사청구일자 2021년01월05일
(65) 공개번호 10-2022-0098991
(43) 공개일자 2022년07월12일
(56) 선행기술조사문헌
Bakhshi, Ali, Aaron SW Wong, and Stephan Chalup. "End-to-end speech emotion recognition based on time and frequency information using deep neural networks." ECAI 2020. IOS Press, 2020. 969-975.*
Kurpukdee,Nattapong et.al, Speech emotion recognition using convolutional long short-term memory neural network and support vector machines, APSIPA ASC, 2017, IEEE, Dec. 2017, Vol.2017, no.12, pp.1744-1749*
Mustaqeem, et.al, CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network, Mathematics, MDPI AG, Dec. 2020, Vol.8, no.12, pp.2133*
이상현, 김재동, 고한석, 강인한 감정 특징 추출을 위한 End-to-end 기반의CRNN-GLU-ATT 모델, 전자공학회논문지(2020, vol.57, no.10, pp. 45-55 (11 pages), Oct. 2020*
*는 심사관에 의하여 인용된 문헌

(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 2 항 심사관 : 김영신

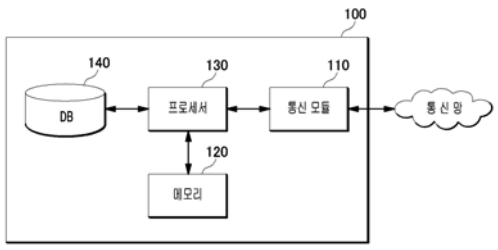
(54) 발명의 명칭 **음성 신호에 기반한 감정 인식 장치 및 방법**

(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 포함하며, 상기 음성 기반 감정 인식 프로그램은, 발화자의 음성 데이터를 수신하고, 수신한 음성 데이터를 감정 분류 모델에 입력하

(뒷면에 계속)

대 표 도 - 도1



여 발화자의 감정을 분류한다. 이때, 상기 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하고, 상기 로컬 특징과 글로벌 특징에 기반하여 발화자의 감정을 분류한다.

(52) CPC특허분류
G10L 17/02 (2013.01)
G10L 17/04 (2013.01)

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 2020.06.01 ~ 2021.02.28
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 데이터를 수신하고, 수신한 음성 데이터를 감정 분류 모델에 입력하여 발화자의 감정을 분류하되,

상기 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하고, 상기 로컬 특징과 글로벌 특징에 기반하여 발화자의 감정을 분류하며,

상기 로컬 특징 추출부는,

복수의 로컬 기능 학습 블록이 순차적으로 연결된 구조를 갖되, 각각의 로컬 기능 학습 블록은 ConvLSTM 계층, BN 계층 및 풀링 계층이 순차적으로 연결된 구조를 가지고,

상기 글로벌 특징 추출부는,

적층된 2개의 GRU(gated recurrent unit)를 각각 포함하는 복수의 단위 레이어를 포함하며,

상기 감정 분류 모델은,

중심 손실 함수와 소프트 맥스 손실함수를 기초로 하는 융합 손실 함수를 통해 상기 로컬 특징 추출부와 상기 글로벌 특징 추출부의 출력에 대한 손실을 산출하고, 상기 손실을 최소화하는 방향으로 가중치 업데이트를 수행하는 것인, 음성 기반 감정 인식 장치.

청구항 2

삭제

청구항 3

삭제

청구항 4

삭제

청구항 5

음성 기반 감정 인식 장치를 이용한 감정 인식 방법에 있어서,

발화자의 음성 데이터를 수신하는 단계, 및

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

상기 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하고, 상기 로컬 특징과 글로벌 특징에 기반하여 발화자의 감정을 분류하고,

상기 로컬 특징 추출부는,

복수의 로컬 기능 학습 블록이 순차적으로 연결된 구조를 갖되, 각각의 로컬 기능 학습 블록은 ConvLSTM 계층, BN 계층 및 풀링 계층이 순차적으로 연결된 구조를 갖고,

상기 글로벌 특징 추출부는,

적층된 2개의 GRU(gated recurrent unit)를 각각 포함하는 복수의 단위 레이어를 포함하며,

상기 감정 분류 모델은,

중심 손실 함수와 소프트 맥스 손실함수를 기초로 하는 융합 손실 함수를 통해 상기 로컬 특징 추출부와 상기 글로벌 특징 추출부의 출력에 대한 손실을 산출하고, 상기 손실을 최소화하는 방향으로 가중치 업데이트를 수행하는 것인, 음성 기반 감정 인식 방법.

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 기계 학습 모델을 통해 음성 신호로부터 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 데이터에서 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, 연구자들은 다양한 유형의 컨볼루션 연산을 통해 숨겨진 신호를 추출하고 선, 곡선, 점, 모양 및 색상을 인식하고 있다. 예를 들면, CNN(convolution neural networks), RNN(recurrent neural networks), LSTM(long short-term memory), DBN(deep belief networks) 등을 포함하는 중간 수준의 종단 간 모델을 활용하고 있다. 다만, 이러한 다양한 인공 신경망 모델의 구성이 여전히 부실하기 때문에 정확도 수준과 인식률이 낮다는 문제가 존재한다. CNN을 이용한 모델의 경우 감정 인식의 정확도를 높이는 역할이 부족하다.

[0005] 또한, 시간에 있어서 장기적인 변화요소를 학습하고, 감정을 인식하기 위해 RNN과 LSTM을 활용하고 있는데, 정확도를 크게 향상시키지 못하면서도 전체 모델의 계산 및 학습 시간을 증가시키는 문제가 있다. 이와 같이, 공간적 감정 신호와 순차적 신호를 인식하는 효율적이고 중요한 프레임 워크를 제공할 필요가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 일본 등록특허공보 제6732703호 (발명의 명칭: 감정 인터렉션 모델 학습 장치, 감정 인식장치, 감정 인터렉션 모델 학습 방법, 감정 인식 방법, 및 프로그램)

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 음성신호의 공간적 특징과 시간적 특징을 모두 활용하여 음성으로부터 발화자의 감정을 분류할 수 있는 음성 기반 감정 인식 장치 및 방법을 제공하는데 목적이 있다.

[0008] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0009] 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 포함하며, 상기 음성 기반 감정 인식 프로그램은, 발화자의 음성 데이터를 수신하고, 수신한 음성 데이터를 감정 분류 모델에 입력하여 발화자의 감정을 분류한다. 이때, 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하고, 상기 로컬 특징과 글로벌 특징에 기반하여 발화자의 감정을 분류한다.

[0010] 또한, 본 발명의 다른 측면에 따른 음성 기반 감정 인식 장치를 이용한 구조물 감정 인식 방법은, 발화자의 음성 데이터를 수신하는 단계, 및 수신한 음성 데이터를 감정 분류 모델에 입력하여 발화자의 감정을 분류하는 단계를 포함하되, 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하고, 상기 로컬 특징과 글로벌 특징에 기반하여 발화자의 감정을 분류를 포함한다.

발명의 효과

[0011] 전술한 본원의 과제 해결 수단에 의하면, 음성 데이터에 포함된 시간적 특징과 공간적 특징을 효과적으로 추출하여, 발화자의 감정을 자동으로 분류할 수 있다.

[0012] 특히, 음성 데이터의 로컬 특징을 추출하는 과정에서 ConvLSTM 모델을 사용함에 따라, 음성 신호의 연속적인 시퀀스를 쉽게 인식하고, 인식된 시퀀스로부터 연결된 감정정보를 추출할 수 있다. 또한, GRU를 통해 서로 시간적으로 떨어져 있는 감정 정보를 함께 고려할 수 있어서, SER 시스템의 예측 성능을 향상시킬 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 장치를 설명하기 위한 개념도이다.

도 3은 본 발명의 일 실시예에 따른 음성 기반 감정 인식 방법을 설명하기 위한 순서도이다.

도 4는 본 발명의 일 실시예에 따른 음성 기반 감정 인식 방법에 사용되는 감정 분류 모델의 구축 과정을 설명하기 위한 순서도이다.

도 5는 본 발명의 일 실시예에 따른 ConvLSTM 계층의 구체적인 구성을 도시한 도면이다.

도 6은 본 발명의 일 실시예에 따른 글로벌 특징 추출부에 사용되는 GRU의 구성을 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 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 데이터를 생성하는 것도 가능하다.

[0020] 통신 모듈(110)은 발화자의 음성 데이터를 외부 기기로부터 수신하는 것으로서, 각종 스마트 단말에 연결된 마이크 등을 통해 입력된 음성 데이터를 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.

[0021] 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.

[0022] 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.

[0023] 이때, 음성 기반 감정 인식 프로그램은, 발화자의 음성 데이터를 통신 모듈(110)을 통해 수신하고, 수신한 음성 데이터를 감정 분류 모델에 입력하여 발화자의 감정을 분류한다. 이때, 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하며, 추가적으로 손실 함수를 통해 감정 분류 모델을 갱신하는 손실 함수부를 포함할 수 있다. 음성 기반 감정 인식 프로그램의 구체적인 내용에 대해서는 추후 설명하기로 한다.

[0024] 이때, 메모리(120)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치를 통칭하는 것이다.

[0025] 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.

[0026] 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.

[0027] 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.

[0028] 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.

[0029] 한편, 음성 기반 감정 인식 장치(100)는 장치(100)에 내장되거나 이에 접속된 마이크 등을 통해 발화자의 음성 신호를 녹음하여, 음성 데이터를 직접 생성할 수 있으며, 이에 대해 감정 인식을 수행할 수 있다.

[0030] 도 2는 본 발명의 일 실시예에 따른 음성 기반 감정 인식 장치를 설명하기 위한 개념도이고, 도 3은 본 발명의 일 실시예에 따른 음성 기반 감정 인식 방법을 설명하기 위한 순서도이고, 도 4는 본 발명의 일 실시예에 따른 음성 기반 감정 인식 방법에 사용되는 감정 분류 모델의 구축 과정을 설명하기 위한 순서도이다.

[0031] 메모리(140)에 저장된 음성 기반 감정 인식 프로그램에 의해 수행되는 음성 기반 감정 인식 방법을 살펴보기로 한다.

[0032] 먼저, 음성 기반 감정 인식 장치(100)에 설치된 음성 기반 감정 인식 프로그램은 마이크 등을 통해 기록된 음성 데이터를 마이크로부터 수신하거나, 통신 모듈(110)을 통해 음성 데이터를 수신한다(S310). 음성 데이터는 디지털 데이터로서, 소정의 시간 단위로 구분된 음성 세그먼트로 분리되어, 감정 분류 모델에 입력될 수 있다. 이와 같이, 서로 연속된 관계에 있는 음성 세그먼트들은 시간적으로 강한 상관 관계를 갖게되며, 이러한 특징을 이용하여 감정 인식을 수행한다.

[0033] 다음으로, 음성 기반 감정 인식 프로그램은 수신한 음성 데이터를 감정 분류 모델에 입력하여 발화자의 감정을 분류한다(S320). 예를 들면, 음성 데이터를 감정 분류 모델에 입력함에 따라, 그 출력으로서 발화자의 감정 상태를 '화남', '슬픔', '행복', '보통' 등으로 분류할 수 있다.

[0034] 이때, 감정 분류 모델은 ConvLSTM을 통해 음성 데이터의 로컬 특징을 추출하는 로컬 특징 추출부, GRU(gated Recurren Unit)를 통해 음성 데이터의 글로벌 특징을 추출하는 글로벌 특징 추출부를 포함하는 것으로, 이의 구체적인 구성과 구축 과정에 대해서는 도 2, 도 4 내지 도 6을 통해 더욱 상세히 살펴보기로 한다.

[0035] 본 발명에서 처리하는 음성 데이터는 시간적으로 연속되는 특징을 가진 데이터로서, 통상적으로는 LSTM과 같은 모델을 사용하여 특징을 추출하고 있으나, 해당 모델의 경우 계산 및 학습 시간을 증가시키는 문제점이 있다.

[0036] 이에, 본 발명의 감정분류 모델은 공간적인 특징을 추출하고 학습하는데 유용한 CNN 모델과 시간적인 특징을 추출하고 학습하는데 유용한 LSTM 모델을 병합한, ConvLSTM 을 사용하는 복수의 로컬 기능 학습 블록 (LFLB, local features learning blocks)들로 이루어진 로컬 특징 추출부를 포함한다. 본 발명에서는 공간적 특징의 활용함으로써, 음성 세그먼트들 간의 시간적 간격 뿐만 아니라 주파수 대역에서 간격을 두고 분리되어 있는 특징을 추출하여 감정을 분류하는데 활용한다. 이를 통해 짧은 시간뿐만 아니라 긴 시간에 걸쳐 표현되는 감정의 특징을 적절히 활용할 수 있게 해주어 감정인식의 성능향상에 이바지한다.

[0037] 도면에서는, 4개의 로컬 기능 학습 블록 (LFLB)이 순차적으로 연결된 구조를 제시하고 있는데, 이는 예시적인 구성으로서 본 발명이 이에 제한되는 것인 아니다.

[0038] 이때, 각각의 로컬 기능 학습 블록은 도 2에서와 같이, ConvLSTM 계층, BN 계층 및 풀링 계층이 순차적으로 연결된 구조를 가진다. 그리고, 각각의 복수의 로컬 기능 학습 블록이 순차적으로 연결된 구조를 통해, 음성 세그먼트 간의 입력-상태(input-state) 및 상태-상태(state-state) 상관 관계를 찾을 수 있다. 즉, 순차적으로 입력된 음성 세그먼트를 처리하는 과정에서 각 음성 세그먼트의 상관 관계를 포착하고, 이를 통해 감정을 인식한다.

[0039] ConvLSTM 계층은 시퀀스를 최적화하고 음성 세그먼트 간의 시공간적 상관 관계를 찾기 위해, 순차적 정보를 내부 상태로 유지하기 위해 숨겨진 단계별 예측에 사용되었다.

[0040] 도 4를 참조하여, 감정 분류 모델의 구축 과정을 살펴보기로 한다.

[0041] 먼저, 음성 기반 감정 인식 장치(100)에 설치된 음성 기반 감정 인식 프로그램은 마이크 등을 통해 기록된 음성 데이터를 마이크로부터 수신하거나, 통신 모듈(110)을 통해 음성 데이터를 수신한다(S410).

[0042] 다음으로, ConvLSTM 에 기반하여 로컬 특징 추출부에 음성 데이터를 입력한다(S420).

[0043] 도 5는 본 발명의 일 실시예에 따른 ConvLSTM 계층의 구체적인 구성을 도시한 도면이다.

[0044] 도시된, ConvLSTM 계층은 다음의 수학식을 이용하여 가중치를 계산한다.

[0045] [수학식 1]

$$i_t = \sigma\left(w_{ix} * x_t + w_{ih} * h_{t-1} + w_{ic} \odot c_{t-1} + b_i\right)$$
$$f_t = \sigma\left(w_{fx} * x_t + w_{fh} * h_{t-1} + w_{fc} \odot c_{t-1} + b_f\right)$$
$$o_t = \sigma\left(w_{ox} * x_t + w_{oh} * h_{t-1} + w_{oc} \odot c_t + b_o\right)$$
$$g_t = tanh\left(w_{gx} * x_t + w_{gh} * h_{t-1} + b_g\right)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh\left(c_t\right)$$

[0046]

[0047] σ 는 시그모이드 함수를 나타내고, * 는 컨볼루션 연산을 나타내고, ⓒ는 엘리먼트별 연산(element wise operation), tanh는 쌍곡탄젠트 함수(hyperbolic tangent function), w는 각각의 변수에 대한 가중치, b는 편향 값, t는 연산 반복 횟수, x_t 는 입력 데이터, c_t는 셀 상태(cell state), h_t는 은닉 상태(hidden state)를 나타낸다. 이와 같이, ConvLSTM 계층에서는 행렬간의 곱이 행해지던 연산의 일부가 컨볼루션 연산으로 대체된다.

[0048] 그리고, 도 5에서 i_t는 입력 게이트(input gate), f_t 는 망각 게이트(forget gate), o_t는 출력 게이트(output gate), g_t는 입력 변조 게이트(input modulation gate)를 각각 나타내며, 이는 일반적인 LSTM의 구성과 동일하다.

[0049] 한편, ConvLSTM에서는 각 입력 게이트에서 처리되는 데이터와 입력 데이터(x_t), 셀 상태(c_t), 은닉 상태(h_t) 는 모두 3차원 텐서로 표현된다. 이때, 입력 텐서에서 첫 번째 차원은 시간 정보, 두 번째 차원은 크기 정보, 세번째 차원은 공간 정보를 나타낸다. 이와 같이, ConvLSTM은 상태에서 상태로 전환하는 동안 시공간 특징을 추출하는 것에 기술적 특징이 있다.

[0050] 다시 도 4를 참조하면, GRU 기반의 글로벌 특징 추출부에 음성 데이터를 입력한다(S430).

[0051] 먼저, 도 2에 도시된 바와 같이, 글로벌 특징 추출부는 GFLB(Global Feature Learning Block)를 포함한다. GFLB는 음성 데이터에서 글로벌 특징 정보를 학습하고, 장기적인 컨텍스트 종속성을 인식하기 위해 GRU(gated recurrent unit)를 포함한다.

[0052] 도 6은 본 발명의 일 실시예에 따른 글로벌 특징 추출부에 사용되는 GRU의 구성을 도시한 것이다.

[0053] GRU는 게이트 메커니즘이 적용된 LSTM 프레임워크의 일종으로서, (a)에 도시된 바와 같이, 업데이트 게이트 및 리셋 게이트를 포함한다. 업데이트 게이트는 LSTM에서의 망각 게이트 및 입력 게이트와 같은 동작을 수행하고, 리셋 게이트는 LSTM에서의 리셋 게이트와 같은 동작을 수행한다.

[0054] 리셋 게이트는 과거의 정보를 적당히 리셋시키는 것으로서, 시그모이드 함수를 이용하며, 아래 수학식과 같이 r_t^j 를 출력한다.

[0055] [수학식 2]

$$r_t^j = \sigma\left(W_r x_t + r^{h_{t-1}}\right)^j$$

[0056]

[0057] 업데이트 게이트는 과거와 현재의 정보의 최신화 비율을 결정하는 것으로, 시그모이드 함수를 이용하며, 아래 수학식과 같이 z_t^j 를 출력한다.

[0058] [수학식 3]

[0059] $$z_t^j = \sigma\left(W_x x_t + U_z h_{t-1}\right)^j$$

[0060] 또한, 업데이트 게이트는 수학식 4를 통해 현시점의 정보 후보군($\hat{h}_t^j$)을 산출하는데, 이때 리셋 게이트의 결과를 이용한다.

[0061] [수학식 4]

[0062] $$\hat{h}_t^j = \tanh\left(W x_t + U(r_t . * h_{t-1})\right)^j$$

[0063] 마지막으로, 최종 은닉 상태(h_t^j)의 결과는 수학식 3과 수학식 4에 의해서 결정되는 업데이트 게이트의 출력을 결합하여 수학식 5에 의해 결정된다.

[0064] [수학식 5]

[0065] $$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j \hat{h}_t^j$$

[0066] 이때, σ 는 시그모이드 함수를 나타내고, * 는 엘리먼트별 곱셈(element wise multiplication), tanh는 쌍곡탄젠트 함수(hyperbolic tangent function), W와 U는 각 변수에 대한 가중치, t는 연산 반복 횟수, x_t 는 입력 데이터, h_t는 은닉 상태를 나타낸다.

[0067] 이와 같은 구성에 의해, 음성 세그먼트에 포함된 단기 종속성은 리셋 게이트에 의해 활성화되고, 음성 세그먼트의 이전 상태는 업데이트 게이트에 의해 제어되는데, 업데이트 게이트는 장기적인 상황 정보를 제어하는 역할도 수행한다.

[0068] 한편, 본 발명에서는 (b)에 도시된 바와 같이, 2개의 GRU를 적층(stack)한 단위 레이어를 복수개 배치하여 글로벌 특징에 대한 가중치를 조절할 수 있다.

[0069] 그리고, 글로벌 특징 추출부의 출력단에는 완전 연결된(fully connected) 레이어가 결합되며, 이를 통해 발화자의 감정을 분류하며, 이후에 결합되는 융합 손실 함수의 결과를 기초로 갱신될 수 있다.

[0070] 다시 도 4를 참조하면, 손실 함수를 이용하여 감정 분류 모델을 갱신하는 작업을 수행한다(S440).

[0071] 본 발명에서는 중심 손실 함수(center loss function)와 소프트 맥스 손실 함수를 사용하여 감정 분류 모델의 손실을 산출한다. 소프트 맥스 손실 함수를 이용한 모델의 예측 성능은 클래스 내에서 거리가 멀기 때문에 다소 성능이 낮아진다.

[0072] 본 발명에서는 중심 손실 함수를 사용하여 클래스 내 최소 거리를 계산하고 소프트 맥스 손실 함수를 통해 클래스 간 최대 거리를 계산하였으며, 구체적인 수학식은 아래와 같다.

[0073] [수학식 6]:소프트 맥스 손실 함수

[0074] $$L_s = -\sum_{i=1}^{m} \log \frac{e^{w_{yi}^T \cdot x_i + b_{yi}}}{\sum_{j=1}^{n} e^{w_{yi}^T \cdot x_i + b_{yi}}}$$

[0075] [수학식 7]: 중심 손실 함수

[0076] $$L_c = \frac{1}{2}\sum_{i=1}^{m} ||x_i - c_{yi}||_2^2$$

[0077] n은 클래스의 개수, m은 최소 배치 사이즈, c_{yi}는 클래스y_i의 중심을 나타낸다.

[0078] 이때, 실시간 시나리오에서 오 분류를 방지하는 데 필요한 최소 거리를 계산하기 위해 중심 손실에 대한 λ 기호를 사용하여, 소프트 맥스 손실 함수와 중심 손실 함수를 모두 반영한, 융합 손실 함수를 수학식 8과 같이 사용하였다.

[0079] [수학식 8]

[0080] $$L = L_S + \lambda L_c$$

[0081] 감정 분류 모델은 중심 손실 함수와 소프트 맥스 손실함수를 기초로하는 융합 손실 함수를 통해 로컬 특징 추출부와 글로벌 특징 추출부의 출력에 대한 손실을 산출하고, 손실을 최소화하는 방향으로 가중치 업데이트를 수행한다.

[0082] 이와 같이 구성된 본 발명의 감정 분류 모델의 효과를 평가하기 위해 동 분야에서 학문적 실험에 널리 사용되는 오픈 데이터베이스인 IEMOCAP 및 RAVDESS를 사용하였는데, 이들은 각각 감정적 언어 말뭉치를 포함하는 두 가지 표준 말뭉치 데이터를 포함한다. 본 발명에 따른 IEMOCAP와 RAVDESS 말뭉치에 대해 각각 75 %의 인식률과 80 %의 인식률을 확보하였으며, 이는 2020년 말 기준으로 최상위의 수치에 해당한다.

[0083] 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.

[0084] 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.

[0085] 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.

[0086] 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

[0087] 100: 음성 기반 감정 인식 장치

110: 통신 모듈

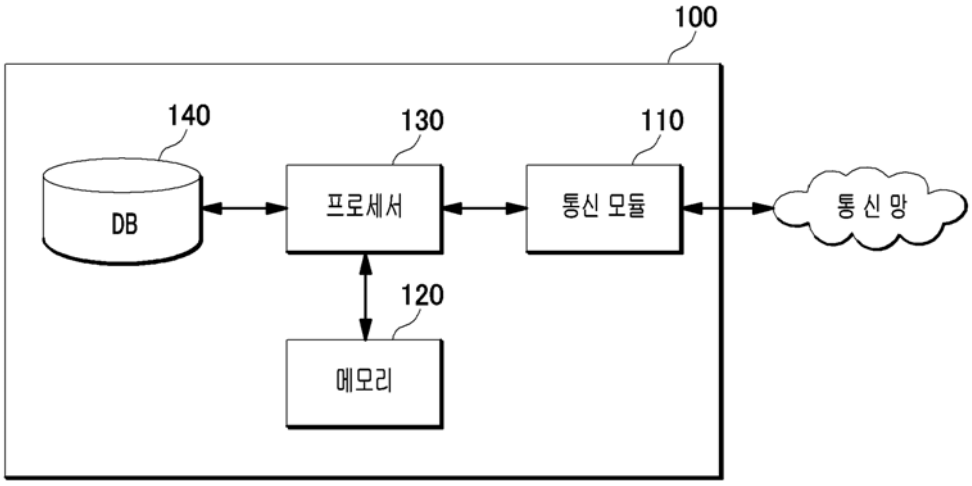
120: 메모리

130: 프로세서

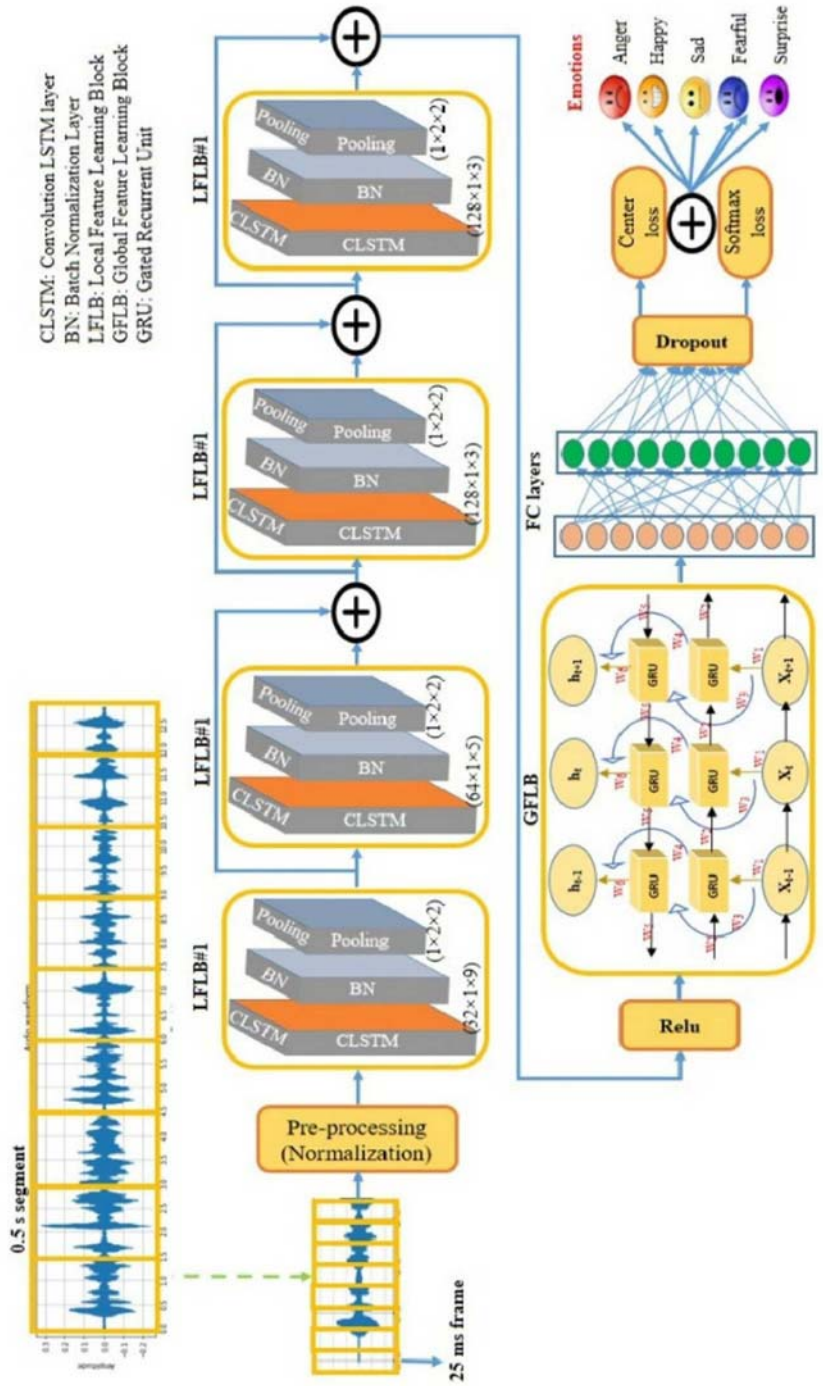
140: 데이터베이스

도면

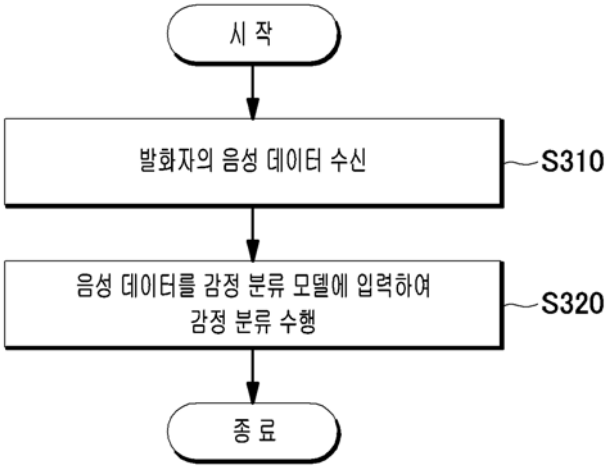
도면1



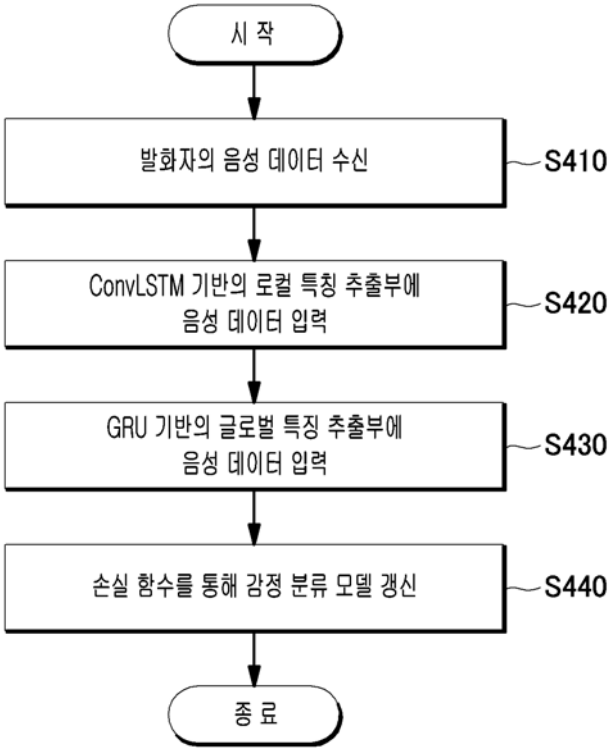
도면2


CLSTM: Convolution LSTM layer
BN: Batch Normalization Layer
LFLB: Local Feature Learning Block
GFLB: Global Feature Learning Block
GRU: Gated Recurrent Unit
0.5 s segment
25 ms frame
Pre-processing (Normalization)
LFLB#1
Pooling
BN
CLSTM
(1×2×2)
(128×1×3)
(64×1×5)
(32×1×9)
GFLB
GRU
FC layers
Dropout
Relu
Center loss
Softmax loss
Emotions
Anger
Happy
Sad
Fearful
Surprise

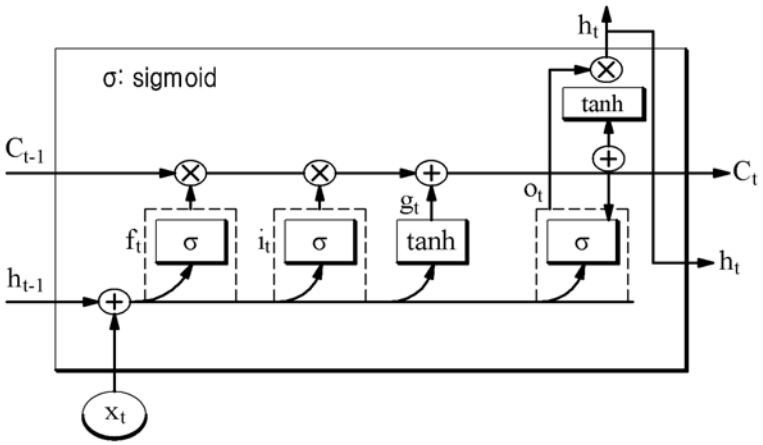
도면3


시 작
발화자의 음성 데이터 수신
S310
음성 데이터를 감정 분류 모델에 입력하여
감정 분류 수행
S320
종 료

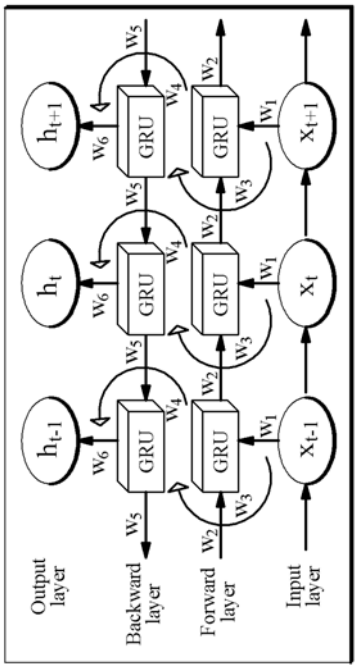
도면4


시 작
발화자의 음성 데이터 수신
S410
ConvLSTM 기반의 로컬 특징 추출부에
음성 데이터 입력
S420
GRU 기반의 글로벌 특징 추출부에
음성 데이터 입력
S430
손실 함수를 통해 감정 분류 모델 갱신
S440
종 료

도면5

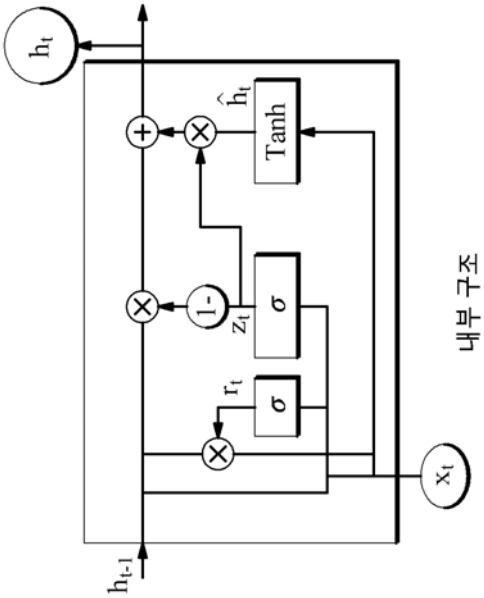


도면6



외부 구조

(b)



내부 구조

(a)