(19) 대한민국 지식재산처(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2026년05월28일
(11) 등록번호 10-2971065
(24) 등록일자 2026년05월26일

(51) 국제특허분류(Int. Cl.)
G10L 25/63 (2013.01) *G06N 3/0464* (2023.01)
G10L 19/26 (2013.01) *G10L 21/10* (2013.01)
G10L 25/18 (2013.01) *G10L 25/30* (2013.01)
(52) CPC특허분류
G10L 25/63 (2013.01)
G06N 3/0464 (2023.01)
(21) 출원번호 10-2024-0197776
(22) 출원일자 2024년12월26일
심사청구일자 2024년12월26일
(56) 선행기술조사문헌
Slimi, Anwer, Henri Nicolas, and Mounir Zrigui, Hybrid Time Distributed CNN-transformer for Speech Emotion Recognition, ICSOFT, 2022*
Mustaqeem et.al, A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition, Sensors, MDPI, Dec. 2019, Vol.20, no.1, pp.183*
Madan,Anjum et.al, CNN-Based Models for Emotion and Sentiment Analysis Using Speech Data, ACM transactions on Asian and low-resource language information processing, Association for Computing Machine*
Ristea, Nicolae-Catalin, Radu Tudor Ionescu, and Fahad Shahbaz Khan, Septr: Separable transformer for audio spectrogram processing, arXiv preprint arXiv:2203.09581, 2022*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
국립한밭대학교 산학협력단
대전광역시 유성구 동서대로 125 (덕명동)
(72) 발명자
이승호
대전광역시 유성구 엑스포로123번길 26-30 스마트시티리버뷰오피스텔 102동 601호
김정윤
대전광역시 유성구 덕명로98번길 5-6, 101호
(74) 대리인
이은철, 이우영

전체 청구항 수 : 총 9 항 심사관 : 김영신

(54) 발명의 명칭 **음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템 및 그 방법**

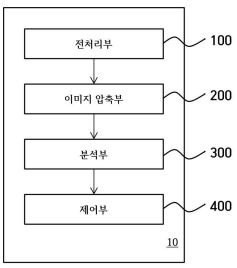
(57) 요 약

본 발명은 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템에 관한 것이며, 음성 감정 인식 정확도를 향상시키는 시스템에 관한 것이다.

본 발명에 따르면, 입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 전처리부, 변환된 log-Mel 스펙트로그

(뒷면에 계속)

대 표 도



램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하는 이미지 압축부, 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 Vision Transformer 기반 분석부 및 상기 분석 결과, 감정 상태 결과를 출력하는 제어부를 포함하여 CNN기반의 이미지압축을 통해 음성 spectrogram에서 중요한 주파수 영역을 시간(수평)적으로 압축하여 vision transformer의 음성 감정인식정확도를 향상시키는 효과가 있다.

(52) CPC특허분류

G10L 19/26 (2013.01)
G10L 21/10 (2013.01)
G10L 25/18 (2013.01)
G10L 25/30 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2710013554
과제번호	2022R1F1A1066371
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	메타버스 컨텐츠에서 사용할 수 있는 음성에 따른 자연스러운 감정 표현이 가능한 3D 얼굴 자동 생성 솔루션 개발
과제수행기관명	국립한밭대학교
연구기간	2024.03.01 ~ 2025.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1345370812
과제번호	2021RIS-004
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	지자체대학협력기반지역혁신사업
연구과제명	(대전세종충남지역혁신플랫폼)충남대학교
과제수행기관명	충남대학교
연구기간	2024.03.01 ~ 2025.02.28

명 세 서

청구범위

청구항 1

입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 전처리부(100),

변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하되, 상기 수평 방향 커널의 크기 및 stride를 조절하여 억양 변화, 강도 변화, 감정에 따른 음소 길이와 같은 감정 관련 시간 프레임 특징을 강조하는 이미지 압축 연산 프로세서를 포함하는 이미지 압축부(200),

상기 이미지 압축부(200)에서 출력된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 수행하는 Vision Transformer 기반 연산 모듈을 포함하고, 감정 상태를 분석하는 분석부(300) 및

상기 분석 결과를 바탕으로 감정 상태를 출력하고 시스템 외부로 전달하는 제어 프로세서를 포함하는 제어부(400)를 포함하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 2

제1항에 있어서,

상기 전처리부는 입력 음성 신호를 받아 log-Mel 스펙트로그램으로 변환하되, 상기 음성 신호의 시간-주파수 특성을 추출하기 위해 Short-Time Fourier Transform(STFT)을 수행하고, Mel 필터뱅크와 로그 변환을 통해 log-Mel 스펙트로그램을 생성하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 3

제1항에 있어서,

상기 이미지 압축부는

log-Mel 스펙트로그램의 시간 축에 대해 수평 방향 커널을 적용하는 CNN 구조로 이루어지며,

상기 수평 방향 커널의 크기 및 stride를 조절함으로써 시간 프레임 단위로 이동하면서 감정 상태 분류에 유의미한 억양 변화, 강도 변화, 주파수 대역의 에너지 분포, 감정에 따른 음소 길이 및 시간 프레임 내 주파수 축 방향의 에너지 분포 패턴을 강조하도록 구성되고,

시간 프레임 내의 에너지 분포는 감정의 억양, 강도에 따라 특정 주파수 대역에 에너지가 집중되는 공간적 패턴으로 나타나는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 4

제1항 또는 제3항에 있어서,

상기 이미지 압축부에서, CNN은 3x5 및 3x7 크기의 수평 커널과 stride 값을 조절하여 스펙트로그램 이미지를 점진적으로 압축하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 5

제1항 또는 제3항에 있어서,

상기 이미지 압축부에서, CNN은 log-Mel 스펙트로그램의 불필요한 시간-주파수 성분을 제거하여 입력 이미지의 크기를 128x1001에서 128x129로 축소하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 6

제1항에 있어서,

상기 분석부는

CNN 기반 이미지 압축부에서 생성된 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용하여 패치 단위의 전역적 및 지역적 시간-주파수 특징을 분석하고,

분석된 정보를 바탕으로 음성 감정 상태를 분류하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 7

제1항에 있어서,

상기 분석부는 6개의 어텐션 헤드와 5층 깊이의 Multi-Head Attention 구조를 포함하여 입력 데이터를 분석하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 8

제1항에 있어서,

상기 제어부는 분석부에서 도출된 감정 상태 결과를 "기쁨", "슬픔", "분노"의 카테고리로 출력하되, 감정 상태 결과를 시각적 또는 텍스트 형식으로 제공하는 사용자 인터페이스를 포함하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템.

청구항 9

음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템을 이용한 방법에 있어서,

(a) 상기 시스템이 입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 단계

(b) 상기 시스템이 변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하되, 커널의 크기 및 stride를 조절하여 억양 변화, 강도 변화, 감정에 따른 음소 길이와 같은 감정 관련 시간 프레임 특징을 강조하는 단계 및

(c) 상기 시스템이 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 단계를 포함하는 것을 특징으로 하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템을 이용한 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템에 관한 것이며, 보다 상세하게는 Log-Mel 스펙트로그램을 수평(시간)적으로만 압축하여 음성 감정 인식 정확도를 향상시키는 시스템에 관한 것이다.

배 경 기 술

[0002] 음성 감정 인식(Speech Emotion Recognition, SER)은 인간의 감정을 음성 신호를 통해 자동으로 인식하는 기술로, 감정 분석, 인간-컴퓨터 상호작용(HCI), 고객 서비스 시스템, 건강 관리 등 다양한 분야에서 중요한 역할을 하고 있다.

[0003] 감정은 음성 신호의 억양, 강세, 속도, 톤 등 다양한 특성에 반영되며, 이를 정확하게 추출하고 해석하는 것이 SER의 핵심 과제이다. 최근 SER 분야는 딥러닝 기반의 접근법을 통해 뛰어난 성과를 얻고 있으며, 특히 CNN(Convolutional Neural Networks)과 transformer 모델들이 뛰어난 성능을 보이고 있다.

[0004] 그러나 음성 감정 인식에서 중요한 도전 과제 중 하나는 음성 데이터의 복잡한 비선형적 특징과 환경적 요인(예: 배경 소음, 화자의 발음 차이 등)에 의해 생성되는 어텐션 노이즈 문제이다. 어텐션 메커니즘은 중요한 정보를 강조하는 데 유용하지만, 과도한 노이즈나 불필요한 정보가 강조될 수 있는 단점이 있다.

[0005] 이로 인해 감정 인식 성능이 저하될 수 있다. 이러한 문제를 해결하기 위한 다양한 방법들이 제안되고 있지만, 여전히 음성 감정 인식에서의 정확도 향상과 노이즈 감소는 중요한 연구 과제로 남아 있다.

[0006] 음성신호의 시간-주파수 정보를 추출하는 방법으로 log-Mel 스펙트로그램은 널리 사용된다. log-Mel 스펙트로그램은 음성의 주파수 특성을 잘 포착할 수 있는 특성 덕분에, 최근 많은 음성 인식 및 감정 인식시스템에서 주요한 입력 데이터로 활용되고 있다. 특히 Mel-frequency cepstral 계수(MFCC)나 스펙트로그램 기반의 접근법은 감정의 음향적 특징을 잘 반영하는데 유리하다. 그러나 log-Mel 스펙트로그램은 고차원적이고 희소한 데이터로, 이를 효과적으로 처리하기 위한 모델이 요구된다.

[0007] 최근 연구에서는 이미지 처리 기법을 통해 log-Mel 스펙트로그램을 이미지 형태로 변환하여 시각적 특징을 추출하고, 이를 딥러닝 모델에 입력하는 접근법이 주목받고 있다. 이러한 접근법은 CNN과 같은 모델을 통해 스펙트로그램에서 중요한 시각적 특징을 자동으로 학습할 수 있으며, transformer 계열 모델을 사용한 어텐션 기반의 정교한 분석도 가능하게 한다. 그러나 이미지 압축 및 데이터 해상도의 변화가 모델 성능에 미치는 영향에 대한 연구는 상대적으로 부족하며, 이를 개선하기 위한 방법론이 필요한 상황이다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 특허공개 2015-0087671 (2015.07.30)

발명의 내용

해결하려는 과제

[0009] 본 발명은 CNN기반의 이미지압축을 통해 음성 스펙트로그램에서 중요한 주파수 영역을 시간(수평)적으로 압축하여 vision transformer의 음성 감정인식정확도를 향상시키는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템을 제공하는 데 그 목적이 있다.

과제의 해결 수단

[0010] 본 발명에 따르면, 입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 전처리부(100), 변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하는 이미지 압축부(200), 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 Vision Transformer 기반 분석부(300) 및 상기 분석 결과, 감정 상태 결과를 출력하는 제어부(400)를 포함한다.

[0011] 전처리부는 입력 음성 신호를 받아 log-Mel 스펙트로그램으로 변환하되, 상기 음성 신호의 시간-주파수 특성을 추출하기 위해 Short-Time Fourier Transform(STFT)을 수행하고, Mel 필터뱅크와 로그 변환을 통해 log-Mel 스펙트로그램을 생성한다.

[0012] 이미지 압축부는 log-Mel 스펙트로그램의 시간 축에 대해 수평 방향의 커널을 사용하는 CNN 구조를 적용하여 스펙트로그램 이미지를 압축하고, 노이즈를 제거하며, 시간-주파수 축의 감정 상태를 분류하기 위한 억양 변화, 강도 변화, 주파수 대역의 에너지 분포, 타임-프레임 내 패턴, 감정에 따른 음소 길이 특징을 강조한다. 여기서, 음소 길이 특징을 강조한다는 것은 감정 상태에 따라 달라질 수 있는 음소의 길이 변화를 효과적으로 추출하고 이를 분석하여 감정 상태를 더 정확히 예측할 수 있도록 한다.

[0013] 이미지 압축부에서, CNN은 3x5 및 3x7 크기의 수평 커널과 stride 값을 조절하여 스펙트로그램 이미지를 점진적으로 압축한다.

[0014] 이미지 압축부에서, CNN은 log-Mel 스펙트로그램의 불필요한 시간-주파수 성분을 제거하여 입력 이미지의 크기를 128x1001에서 128x129로 축소한다.

[0015] 분석부는 CNN 기반 이미지 압축부에서 생성된 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘

을 이용하여 패치 단위의 전역적 및 지역적 시간-주파수 특징을 분석하고, 분석된 정보를 바탕으로 음성 감정 상태를 분류한다.

[0016] 분석부는 6개의 어텐션 헤드와 5층 깊이의 Multi-Head Attention 구조를 포함하여 입력 데이터를 분석한다.

[0017] 제어부는 분석부에서 도출된 감정 상태 결과를 "기쁨", "슬픔", "분노" 등의 카테고리로 출력하되, 감정 상태 결과를 시각적 또는 텍스트 형식으로 제공하는 사용자 인터페이스를 포함한다.

[0018] 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템을 이용한 방법에 있어서, (a) 상기 시스템이 입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 단계, (b) 상기 시스템이 변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하는 단계 및 (c) 상기 시스템이 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 단계를 포함한다.

발명의 효과

[0019] 본 발명에 따르면, CNN기반의 이미지압축을 통해 음성 spectrogram에서 중요한 주파수 영역을 시간(수평)적으로 압축하여 vision transformer의 음성 감정인식정확도를 향상시키는 효과가 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 구성도이다.

도 2는 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 개요를 나타낸 것이다.

도 3은 본 실시예에 따른 log-Mel 스펙트로그램 생성 과정을 나타낸다.

도 4는 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 CNN기반 이미지 압축 과정을 나타낸 것이다.

도 5는 본 발명에서 감정 분류에 사용한 vision transformer의 구조를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템에서 제안하는 음성감정인식에서의 어텐션 노이즈감소를 위한 CNN 기반의 Log-Mel 스펙트로그램 이미지 압축기법의 독창성은 다음과 같다.

[0022] 수평 커널을 이용한 CNN 기반 이미지 압축으로 log-Mel 스펙트로그램의 128x1001 크기 이미지를 128x129 크기로 변환하여 임의의 이미지 보간법이 자동으로 수행되도록 설계되었다. 이는 시간(수평)축에 대해서만 주파수 정보를 압축하여 학습된 CNN이 중요하지 않은 시간의 정보는 희박하고 중요한 정보는 선명하게 도출하도록 조절된다.

[0023] 압축된 log-Mel 스펙트로그램 이미지에서 vision transformer를 이용하여 음성 감정 인식을 수행한다. 이를 통해 압축된 이미지에서 중요한 시간축의 정보가 강조되어 더 정확한 어텐션이 수행되며 음성 감정 인식의 정확도가 향상된다.

[0024] 이하, 첨부된 도면을 참조하여 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템을 설명하기로 한다.

[0025] 도 1은 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 구성도이다.

[0026] 도 1에 도시된 바와 같이, 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템(10)은 전처리부(100), 이미지 압축부(200), 분석부(300), 제어부(400)를 포함한다.

[0027] 전처리부(100)는 입력 음성 신호를 log-Mel 스펙트로그램으로 변환한다.

[0028] 전처리부는 입력 음성 신호를 받아 log-Mel 스펙트로그램으로 변환하되, 상기 음성 신호의 시간-주파수 특성을

추출하기 위해 Short-Time Fourier Transform(STFT)을 수행하고, Mel 필터뱅크와 로그 변환을 통해 log-Mel 스펙트로그램을 생성한다. 이러한 전처리부는 음성 신호의 길이를 고정하기 위해 zero-padding을 적용하여 일정한 길이의 스펙트로그램 이미지를 생성한다.

[0029] 참고로, 전처리부(100)는 입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 역할을 수행한다. 예를 들어, 사용자가 "안녕하세요"라는 음성을 입력했다고 가정하면, 전처리부는 다음과 같은 단계를 통해 이를 변환한다.

[0030] 첫째, STFT(Short-Time Fourier Transform)를 수행하여 음성 신호를 작은 시간 구간으로 나눈 후, 각 구간에 대해 주파수 스펙트럼을 계산한다. 이 과정을 통해 음성 신호의 시간-주파수 성분이 포함된 데이터가 생성된다.

[0031] 둘째, Mel 필터뱅크를 적용하여 인간의 청각 특성을 반영한 Mel 스케일로 주파수 데이터를 변환한다. 이는 저주파 영역을 더 세밀하게 처리하고 고주파 영역을 덜 세밀하게 처리하는 방식으로 이루어진다.

[0032] 셋째, 로그 변환을 적용하여 스펙트럼 값의 동적 범위를 줄이고 log-Mel 스펙트로그램을 생성한다.

[0033] 마지막으로, 입력 음성 신호가 길이가 일정하지 않을 경우, zero-padding을 적용하여 모든 데이터를 동일한 길이(예: 4초)로 맞춘다. 이를 통해 짧은 음성 신호의 끝부분에 0을 추가하여 일정한 길이로 변환한다.

[0034] 결과적으로, "안녕하세요"라는 음성 입력은 log-Mel 스펙트로그램으로 변환되어 시간 축(x축)과 주파수 축(y축)의 특징을 나타내는 2D 이미지로 출력된다. 이 이미지는 이후 CNN과 Vision Transformer를 통해 분석되며, 감정 상태를 분류하는 데 사용된다.

[0035] 이미지 압축부(200)는 변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축한다.

[0036] 이러한 이미지 압축부는 log-Mel 스펙트로그램의 시간 축에 대해 수평 방향의 커널을 사용하는 CNN 구조를 적용하여 스펙트로그램 이미지를 압축하고, 노이즈를 제거하며, 시간-주파수 축의 감정 상태를 분류하기 위한 억양 변화, 강도 변화, 주파수 대역의 에너지 분포, 타임-프레임 내 패턴, 감정에 따른 음소 길이 특징을 강조한다.

여기서 '타임-프레임 내 패턴'은 하나의 시간 프레임(고정된 시간 구간) 내에서 주파수 축 방향으로 형성되는 에너지의 공간적 분포를 의미하며, 이는 감정의 억양, 강도에 따라 특정 주파수 대역에서 에너지가 집중되는 패턴으로 나타난다.

[0037] 이미지 압축부에서, CNN은 3x5 및 3x7 크기의 수평 커널과 stride 값을 조절하여 스펙트로그램 이미지를 점진적으로 압축한다.

[0038] 또한 이미지 압축부에서, CNN은 log-Mel 스펙트로그램의 불필요한 시간-주파수 성분을 제거하여 입력 이미지의 크기를 128x1001에서 128x129로 축소한다.

[0039] 참고로, 이미지 압축부(200)는 변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하는 역할을 수행한다. 예를 들어, 128x1001 크기의 log-Mel 스펙트로그램이 주어졌다고 가정하면, CNN은 이를 다음과 같은 방식으로 압축한다.

[0040] 첫째, CNN은 3x5 및 3x7 크기의 수평 커널을 사용하여 시간 축에서 발생하는 패턴 변화를 효과적으로 포착한다. 예를 들어, 특정 시간 구간에서 발생하는 억양 변화나 강도 변화를 강조하고, 불필요한 노이즈는 제거한다.

[0041] 둘째, stride 값을 조절하여 스펙트로그램 이미지를 점진적으로 축소한다. 예를 들어, stride를 2 또는 3으로 설정함으로써 이미지의 크기를 줄이면서도 감정 상태 분류에 중요한 시간-주파수 특징을 보존한다.

[0042] 셋째, 감정 인식에 필요한 주요 특징을 강조한다. 예를 들어, 타임-프레임 내의 억양 변화, 특정 주파수 대역의 에너지 분포, 감정에 따른 음소 길이와 같은 정보를 강조하고, 감정 분석에 불필요한 정보는 걸러낸다.

[0043] 마지막으로, CNN은 log-Mel 스펙트로그램의 크기를 128x1001에서 128x129로 축소한다. 이는 모델이 처리하기 쉬운 크기로 데이터를 압축하면서도 감정 인식에 필요한 핵심 정보를 유지하도록 설계된 과정이다.

[0044] 결과적으로, 이미지 압축부는 log-Mel 스펙트로그램을 효과적으로 압축하여 Vision Transformer로 입력하기 적합한 형태로 변환한다.

[0045] 분석부(300)는 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 Vision Transformer 기반 분석부이다.

[0046] 분석부는 CNN 기반 이미지 압축부에서 생성된 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘

을 이용하여 패치 단위의 전역적 및 지역적 시간-주파수 특징을 분석하고, 분석된 정보를 바탕으로 음성 감정 상태를 분류한다.

[0047] 또한 분석부는 6개의 어텐션 헤드와 5층 깊이의 Multi-Head Attention 구조를 포함하여 입력 데이터를 분석한다.

[0048] 참고로, 분석부(300)는 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 Vision Transformer 기반 분석부이다. 예를 들어, 이미지 압축부를 통해 128x129 크기로 압축된 log-Mel 스펙트로그램이 입력되었을 때, 분석부는 이를 작은 패치 단위로 나누어 각 패치 간의 전역적 및 지역적 시간-주파수 특징을 학습한다.

[0049] 분석부는 Self-Attention 메커니즘을 사용하여 특정 패치가 다른 패치와 얼마나 연관성이 있는지를 계산하고, 중요도가 높은 시간-주파수 특징에 집중한다. 예를 들어, 특정 시간 구간에서 억양 변화나 강도 변화가 발생한 패치는 높은 어텐션 값을 받으며, 감정 상태를 나타내는 주요 특징으로 분석된다.

[0050] 분석부는 6개의 어텐션 헤드와 5층 깊이의 Multi-Head Attention 구조를 사용하여 입력 데이터를 정밀하게 분석한다. 이를 통해, 멀리 떨어진 시간적 패턴 간의 상호작용을 학습하고, 감정 상태 분류에 필요한 전역적 특징과 지역적 특징을 모두 포착한다.

[0051] 결과적으로, 분석부는 입력된 스펙트로그램 이미지에서 도출된 특징을 바탕으로 음성 감정 상태를 "기쁨", "슬픔", "분노" 등으로 분류하는 역할을 수행한다.

[0052] 제어부(400)는 분석 결과, 감정 상태 결과를 출력한다.

[0053] 제어부는 분석부에서 도출된 감정 상태 결과를 "기쁨", "슬픔", "분노" 등의 카테고리로 출력하되, 감정 상태 결과를 시각적 또는 텍스트 형식으로 제공하는 사용자 인터페이스를 포함한다.

[0054] 참고로, 제어부(400)는 분석 결과로 도출된 감정 상태를 "기쁨", "슬픔", "분노" 등의 카테고리로 출력하는 역할을 한다. 예를 들어, 사용자가 "오늘 정말 행복해요!"라는 음성을 입력했을 경우, 분석부가 음성 신호의 특징을 분석하여 "기쁨"이라는 감정 상태를 도출하면, 제어부는 이를 시각적 또는 텍스트 형식으로 사용자에게 제공한다.

[0055] 시각적 형식으로는 화면에 "기쁨"이라는 단어와 함께 관련된 아이콘이나 그래프를 표시할 수 있다. 텍스트 형식으로는 결과를 "현재 감정 상태: 기쁨"과 같은 형태로 출력하여 사용자에게 전달한다.

[0056] 결과적으로, 제어부는 사용자와 시스템 간의 상호작용을 돕는 역할을 하며, 분석된 감정 상태를 직관적이고 이해하기 쉬운 방식으로 제공하여 활용성을 높인다.

[0057] 도 2는 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 개요를 나타낸 것이다.

[0058] 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템(이하, 시스템이라 함)을 이용한 방법에 있어서, (a) 시스템은 입력 음성 신호를 log-Mel 스펙트로그램으로 변환하는 단계, (b) 시스템은 변환된 log-Mel 스펙트로그램을 시간 축에 대해 수평 방향 커널을 사용하는 CNN을 통해 압축하는 단계 및 (c) 시스템은 압축된 스펙트로그램 이미지를 입력받아 Self-Attention 메커니즘을 이용해 감정 상태를 분석하는 단계를 포함한다. (c) 단계 이후, 시스템은 분석부에서 도출된 감정 상태 결과를 "기쁨", "슬픔", "분노" 등의 카테고리로 출력하되, 감정 상태 결과를 시각적 또는 텍스트 형식으로 사용자 인터페이스를 통해 제공하는 단계를 포함한다.

[0059] 본 발명에서 제안하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN 기반의 log-Mel 스펙트로그램 이미지 압축 기법의 개요는 도 2와 같다.

[0060] 첫 번째 단계로 log-Mel 스펙트로그램 생성 및 전처리를 수행하고, 두 번째 단계에서는 CNN 기반 이미지 압축을 수행한다. 마지막으로 vision transformer를 통한 음성 감정 인식을 수행한다.

[0061] 한편, 도 2는 본 발명에서 제안하는 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN 기반의 log-Mel 스펙트로그램 이미지 압축 기법의 전체 과정을 나타낸다.

[0062] Log-Mel 스펙트로그램 생성 및 전처리와 관련하여,

[0063] 본 발명에서는 16kHz의 샘플링 레이트를 가진 음성 신호로부터 log-Mel 스펙트로그램을 생성한다. 음성감정 인

식(SER) 작업에서 효과적으로 음성 신호의 시간-주파수 특성을 포착하기 위해, 다음과 같은 파라미터를 사용한다. hop 크기 64, Hamming window function 길이 512, FFT 포인트 수 1024, 그리고 Mel 밴드수 128. 이러한 설정은 음성 신호의 주파수 해상도를 적절히 반영하며, Mel 스케일이 인간의 청각적 특성을 잘 표현하기 때문에 음성 신호에서 중요한 정보를 잘 추출할 수 있다. 또한, 본 연구에서 사용된 음성 신호의 길이는 대부분 4초 미만이지만, zero padding을 통해 모든 신호의 길이를 4초로 맞춘다. 이를 통해 입력신호의 길이를 고정시키며, 배치 처리를 위한 일관된 입력 크기를 제공한다. 제로 패딩을 적용한 4초 길이의 각 음성신호는 앞서 설정한 파라미터에 의해 128x1001 크기의 log-Mel 스펙트로그램으로 변환된다. 여기서 128은 Mel 주파수 밴드의 수를 나타내며, 1001은 시간 축을 따라 계산된 프레임 수이다.

[0064] Log-Mel 스펙트로그램을 생성하는 과정은 먼저 short-term Fourier transform (STFT)을 적용하여 음성신호의 주파수 스펙트럼을 구한다. 그 후 Mel 필터 뱅크를 통해 주파수 축을 Mel 스케일로 변환하고, 마지막으로 로그 변환을 적용하여 Mel 스펙트로그램의 동적 범위를 압축하고, 음성 신호의 중요한 특성을 강조한다.

[0065] 이 과정에서 사용되는 윈도우 길이 512와 hop 크기 64는 시간 축에서 음성 신호의 중요한 변화를 포착할 수 있는 충분한 해상도를 제공한다. 이와 같은 전처리 과정을 통해 생성된 log-Mel 스펙트로그램은 음성 신호의 중요한 시간-주파수 정보를 잘 보존하며, 이후 모델 학습에 필요한 일관된 입력 크기와 형상을 유지하게 된다. 도 3은 본 실시예에 따른 log-Mel 스펙트로그램 생성 과정을 나타낸다.

[0066] 도 3에 도시된 바와 같이, 음성 신호를 log-Mel 스펙트로그램으로 변환하는 과정을 도식적으로 나타낸 것이다. 음성 신호는 최대 4초의 길이를 가지며, 이 신호는 먼저 Short-Time Fourier Transform(STFT)을 적용받아 시간-주파수 영역의 데이터를 생성한다. 그 후, STFT 결과를 바탕으로 Mel 필터 뱅크를 적용하고, 최종적으로 로그 변환을 통해 log-Mel 스펙트로그램을 생성한다. 생성된 log-Mel 스펙트로그램은 음성 신호의 시간-주파수 정보를 시각적으로 표현하며, 이후의 딥러닝 모델 학습에 필요한 입력 데이터로 활용된다.

[0067] 이 이미지가 시사하는 바는 음성 신호의 시간 및 주파수 정보를 동시에 보존하면서도, 모델 학습에 최적화된 데이터 형식으로 변환하는 전처리 과정의 중요성이다. log-Mel 스펙트로그램은 인간 청각 특성을 반영한 Mel 스케일과 동적 범위를 압축하는 로그 변환을 통해 음성 데이터에서 중요한 특징을 강조한다. 이는 음성 감정 인식과 같은 작업에서 데이터의 품질을 높이고, 모델 성능 향상을 도모할 수 있음을 시사한다.

[0068] 본 발명에서 음성 신호를 log-Mel spectrogram으로 변환하는 작업에 사용된 공식은 [식 1]과 같다.

[0069] [식 1]

[0070] $$STFT(m,k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n-aH] \cdot e^{-j\frac{2\pi}{N}kn}$$

[0071] 여기서, STFT(m.k)는 Short-Time Fourier Transform의 결과값으로, 주어진 시간 구간 (m)과 주파수 (k)에서 음성 신호의 시간-주파수 성분을 나타낸다.

[0072] x[n]은 시간 도메인에서의 입력 신호이다. 이는 샘플링된 음성 신호를 의미한다. w[n-aH]는 윈도우 함수로, 주로 Hamming 윈도우 또는 Hanning 윈도우가 사용된다. n-aH는 특정 시간 구간(aH)을 중심으로 신호를 분리하여 처리하는 역할을 한다. H는 Hop size로, 윈도우를 이동시키는 단위 길이를 나타낸다. a는 윈도우의 이동 스텝 인덱스를 나타낸다. $e^{-j\frac{2\pi}{N}kn}$ 는 주파수 변환을 위한 Fourier 변환으로, N은 Fourier 변환을 수행할 샘플의 총 개수이다. k는 주파수 인덱스를 나타내며, 특정 주파수 성분을 나타낸다. n은 시간 인덱스를 나타낸다.

[0073] 이 공식은 시간-주파수 영역의 정보를 계산하기 위해 Short-Time Fourier Transform(STFT)을 수행하는 수학적 표현이다. x[n]은 입력신호이고, w[n-aH]윈도우 함수는 신호의 특정 구간을 분리하여 해당 구간에 대해 Fourier 변환을 수행할 수 있도록 한다. 이후 $e^{-j\frac{2\pi}{N}kn}$ 를 통해 신호의 주파수 성분을 추출하며, 최종적으로 STFT(m,k)는 시간(m)과 주파수(k)에 대한 성분을 제공한다. 이 공식은 음성 신호를 시간-주파수 영역으로 변환하여 중요한 정보를 보존하는 데 사용된다. log-Mel 스펙트로그램 생성 과정에서 STFT는 초기 단계로, 시간적으로 분할된 각 구간에서 주파수 성분을 추출한다. 이를 통해 음성 신호의 변화 패턴을 분석하고, 이후 Mel 필터 및 로그 변환을 적용하여 인간 청각 특성을 반영한 log-Mel 스펙트로그램을 생성할 수 있다.

[0074] 주어진 입력 이산 신호 ■x[n]과 길이 L의 윈도우 함수 w[n] (이 경우 Hamming), 홉 크기 H, 그리고 이산 푸리에 변환(DFT) 전체 포인트 수 N이 주어진다. STFT(m,k)는 k 번째 frequency bin과 m 번째 시간 프레임에 대한 STFT 계수를 나타낸다. 이후 Mel 스케일로 매핑을 수행하여 log-Mel 스펙트로그램을 생성한다.

[0075] CNN 기반의 이미지 압축과 관련하여,

[0076] 본 실시예에서는 log-Mel 스펙트로그램 이미지를 효과적으로 압축하기 위해 CNN 기반의 이미지 압축 기법을 사용한다. 크기가 (i, j)인 2D 이미지 H와 (k, k) 크기의 2D 커널 F에 대한 CNN의 연산은 [식 2]와 같다.

[0077] [식 2]

[0078] $$G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} H[u,v]F[i-u, j-v]$$

[0079] 여기서, G[i,j]는 CNN의 출력 이미지에서 픽셀 위치(i,j)에 해당하는 값이다. 이는 입력 이미지와 커널 간의 합성곱(Convolution) 연산 결과이다.

[0080] H[u,v]는 CNN 연산에서 사용하는 k*k크기의 커널(필터)이다. 각 필터는 특정 특징(예: 에지, 텍스처 등)을 학습하여 추출하는 역할을 한다.

[0081] F[i-u,j-v]는 입력 이미지의 픽셀 값으로, 커널과 합성곱 연산이 적용되는 영역의 값이다. i-u, j-v는 입력 이미지에서 커널 위치에 따라 참조되는 픽셀 좌표를 나타낸다. 식 2는 커널의 각 위치에 대해 입력 이미지와 커널의 값들을 곱하고 더하는 과정을 나타낸다. 이는 커널의 전체 영역에 대해 연산이 수행됨을 의미한다.

[0082] 이 식은 CNN에서 합성곱 연산(Convolution)을 수학적으로 표현한 것이다. 입력 이미지 F와 필터(커널) H를 합성곱하여 출력 이미지 G를 생성한다. 합성곱 연산은 입력 이미지의 특정 특징을 강조하거나 추출하는 데 사용되며, CNN 기반 모델의 주요 작업 중 하나이다. 합성곱 연산은 다음과 같은 과정을 포함한다:

[0083] 입력 이미지 F의 특정 영역(커널의 크기에 해당)을 선택한다. 선택된 영역의 픽셀 값과 커널 H의 대응하는 값들을 곱한다. 곱한 값들을 모두 더한 후, 결과를 출력 이미지 G의 해당 위치 (i,j)에 저장한다. 이 과정을 이미지 전체에 대해 반복한다.

[0084] 본 실시예에서는 log-Mel 스펙트로그램 이미지를 효과적으로 압축하기 위해 CNN 기반의 합성곱 연산을 활용한다. 합성곱 연산을 통해 입력 이미지의 불필요한 정보를 제거하고, 중요한 특징만을 추출하여 압축한다. 이러한 과정은 데이터의 크기를 줄이는 동시에 의미 있는 정보를 유지하도록 설계되어, 이후 모델 학습과 분석에 적합한 데이터 형식을 제공한다.

[0085] 결론적으로, [식 2]는 log-Mel 스펙트로그램을 효율적으로 처리하고, 음성 감정 인식에서 중요한 특성을 강조하는 CNN 기반 이미지 압축의 핵심 연산을 설명한다.

[0086] 이 방법은 기존의 고차원 log-Mel 스펙트로그램 데이터를 처리하는 데 있어 효율성과 성능을 동시에 고려할 수 있다.

[0087] CNN을 사용한 이미지 압축의 독창성은 수평 커널을 활용하여 스펙트로그램의 해상도를 축소하는 것이다. 수평 커널을 사용하는 이유는 음성 신호의 시간 축에 대한 중요한 변화를 포착하는데 유리하기 때문이다.

[0088] 일반적으로 음성 신호의 감정적 특성은 시간 축에서 발생하는 패턴 변화에 크게 의존하므로, 수평 방향으로의 압축을 통해 중요한 시간적 특성을 유지하면서 해상도를 효율적으로 축소할 수 있다.

[0089] 이 방식은 이미지의 크기를 128x1001에서 128x128로 줄이며 중요한 정보를 추출하는 데 사용된다. CNN을 통한 압축 과정에서, 각 필터는 원본 스펙트로그램의 중요한 특징을 추출하는 역할을 합니다. 여러 층의 컨볼루션을 통해 생성된 특징 맵은 log-Mel 스펙트로그램의 중요한 시간-주파수 정보를 보존하면서, 불필요한 정보를 제거한다. 각 컨볼루션 층은 고차원적인 데이터를 점차적으로 축소하여, 주요 패턴을 추출하고 압축된 이미지에서 중요한 특성만을 강조한다. 이 과정에서 정보 손실을 최소화하고, 모델의 학습 효율성을 높일 수 있다.

[0090] CNN을 사용한 이미지 압축의 또 다른 중요한 목적은 어텐션 메커니즘에서 발생할 수 있는 노이즈를 줄이는 것이다. 원본 log-Mel 스펙트로그램은 고차원적이고 희소한 특성을 가질 수 있기 때문에, 어텐션 메커니즘을 사용하기 전 이 데이터를 압축하는 것은 모델이 중요한 정보를 더 잘 학습할 수 있도록 돕는다.

[0091] 특히, 압축된 이미지에서 불필요한 노이즈를 줄이고, vision transformer 모델이 더 중요한 시간-주파수 특징에 집중할 수 있도록 한다.

[0092] CNN을 사용한 이미지 압축 기법의 가장 큰 장점은 효율적인 정보 축소와 계산 리소스 절감이다. 고차원 데이터를 그대로 사용하기보다는, CNN을 통해 중요한 특징을 추출하고 압축하여, 모델이 학습하는 데 필요한 계산량을 줄일 수 있다.

[0093] 또한, CNN은 파라미터 공유와 지역적 연결을 통해 모델의 파라미터 수를 줄이고, 과적합을 방지하는 데 유리하다. 도 4는 CNN 기반 이미지 압축과정을 나타낸다.

[0094] 수평으로 넓은 3x5 크기의 컨볼루션 커널과 시간 축에 대한 stride 2로 이미지를 1차적으로 압축한다.

[0095] 이후 3x7 크기의 컨볼루션 커널과 시간 축에 대해 stride 3을 적용하여 시간에 대한 주파수 정보를 효과적으로 압축한다.

[0096] 이후 stride 1인 3x5 컨볼루션 커널을 통해 이미지를 깊게 분석한다. 입력 log-Mel 스펙트로그램의 크기는 128x1001이고 압축된 이미지의 크기는 128x129이다.

[0097] 도 4는 본 발명의 일 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 CNN기반 이미지 압축 과정을 나타낸 것이다.

[0098] 도 4에 도시된 바와 같이, 입력 log-Mel 스펙트로그램(128x1001)을 CNN 기반의 이미지 압축 기법을 통해 압축하는 과정을 나타낸다. 이 과정은 여러 층의 컨볼루션 연산(Conv 3x5, Conv 3x7 등)을 활용하며, 각 단계마다 stride와 padding 값을 조정하여 시간-주파수 축의 정보를 점진적으로 압축한다. 최종적으로 압축된 이미지의 크기는 128x129로 줄어들며, 중요한 시간적 특성과 주파수 정보를 보존한다.

[0099] 이러한 도 4와 관련하여, CNN을 활용한 이미지 압축 기법은 고차원 데이터를 다루기 위해 데이터 크기를 줄이면서 중요한 특징을 유지한다. 이는 학습 데이터의 크기를 줄이고 계산 리소스를 절약할 수 있다는 장점을 제공한다.

[0100] 또한 음성 감정 인식에서 시간 축의 변화가 중요한 역할을 하므로, 수평 커널과 다양한 stride 값을 통해 시간-주파수 정보를 효과적으로 보존한다. 이는 원본 데이터의 손실을 최소화하고 특징 추출과 분석에 적합한 입력을 제공한다.

[0101] 고차원 데이터에서 발생할 수 있는 불필요한 정보를 제거함으로써 모델이 더 중요한 특징에 집중할 수 있게 한다. 특히, 압축된 이미지는 어텐션 메커니즘을 사용하는 모델(Vision Transformer)에서 발생할 수 있는 노이즈를 줄여, 인식 정확도를 향상시킨다.

[0102] 효율적인 정보 압축은 학습 속도와 정확도를 높이고, 과적합을 방지하며, 실제 응용에서 계산 자원을 절감한다.

[0103] Vision transformer를 통한 음성 감정 인식과 관련하여,

[0104] 본 발명에서는 vision transformer를 활용하여 음성 감정 인식(SER) 모델을 구축한다. Vision transformer는 최근 컴퓨터 비전 분야에서 큰 주목을 받으며, 이미지 분류 및 패턴 인식에서 뛰어난 성능을 보여온 모델이다. 전통적인 CNN 기반 모델과 달리, vision transformer는 self-attention 메커니즘을 활용하여 입력데이터를 처리하며, 이를 통해 공간적, 시간적 관계를 효과적으로 학습할 수 있다. 본 발명에서는 log-Mel 스펙트로그램을 이미지 형태로 변환하고, 이를 vision transformer에 입력하여 음성 신호에서 감정적 특징을 추출하고 인식하는 방식을 채택하였다.

[0105] 본 실시예에서 제시된 CNN 기반의 이미지 압축 기법을 통해 생성된 128x128 크기의 log-Mel 스펙트로그램 이미지는 vision transformer 모델에 입력된다. Vision transformer는 이미지나 시계열 데이터를 작은 패치(patch)로 나누어 각 패치의 정보를 토대로 전체적인 특징을 학습하는 방식으로 동작한다. Log-Mel 스펙트로그램은 주파수와 시간 축에 따라 나누어진 패치들이 모델에 입력될 수 있도록 변환되며, 이를 통해 음성 신호의 감정적 변화를 모델이 잘 인식할 수 있게 된다.

[0106] Vision transformer의 핵심은 self-attention 메커니즘이다. 이 메커니즘은 입력된 정보에서 중요한 부분을 강조하고, 덜 중요한 부분은 상대적으로 무시하는 방식으로 작동한다. Log-Mel 스펙트로그램의 각 패치는 모델 내에서 attention 값을 계산하여 중요한 주파수와 시간 영역에 집중하게 된다. 이를 통해 모델은 음성신호의 감정적 특징을 효율적으로 추출할 수 있다. 음성 감정 인식에서는 톤, 억양, 속도 등 시간적 특징과 특정 주파수 대

역에서 발생하는 변화를 잘 포착하는 것이 중요하다. Vision transformer는 이처럼 중요한 시간-주파수 영역을 동적으로 선택하여 감정을 인식하는 데 필요한 패턴을 학습한다. ViT는 self-attention 메커니즘을 활용하여 장기 의존성을 학습하는 데 매우 효과적이다. 음성 신호는 짧은 시간 간격 안에서도 감정적인 변화를 나타낼 수 있기 때문에, transformer 모델은 장기적인 시간적 의존성을 모델링하는 데 유리하다. 음성의 억양 변화나 감정의 급격한 전환은 멀리 떨어져 있는 시점들 간의 상호작용을 통해 발생할 수 있다. Vision transformer는 이러한 장기적 의존성을 자연스럽게 학습하며, 보다 정교한 감정 인식이 가능하다. [식 3]은 Attention 메커니즘을 나타낸다.

[0107] [식 3]

$$Attention(Q,K,V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}})V$$

[0108]

[0109] 여기서, Q(Query)는 현재 데이터 포인트의 특징을 나타내며, 특정 패치가 다른 패치와 얼마나 연관되는지를 측정하기 위해 사용된다. K(key)는 모든 데이터 포인트의 특징을 요약한 값으로, Query와의 유사성을 계산하는 데 활용된다. V(Value)는 입력 데이터의 실제 값을 포함하며, Query와 Key의 관계를 통해 계산된 중요도에 따라 가중합으로 출력된다. d_k는 Key의 차원으로, 모델의 안정적인 학습을 위해 정규화에 사용된다. softmax는 계산된 유사도를 확률 분포로 변환하여, 특정 입력이 얼마나 중요한지 강조한다.

[0110] Vision Transformer와 음성 감정 인식에 있어서, Log-Mel 스펙트로그램의 패치 분할에서, 128x128 크기의 log-Mel 스펙트로그램 이미지는 작은 패치로 나뉘어 Vision Transformer에 입력된다. 각 패치는 음성의 주파수 및 시간 정보를 캡처하며, Transformer는 이 패치 간의 관계를 학습한다.

[0111] Self-Attention 메커니즘에서, Self-Attention은 시간적, 주파수적 의존성을 강조하여 음성 신호에서 중요한 부분(예: 특정 억양 변화나 주파수 대역)을 강조한다. ViT는 이러한 메커니즘을 통해 시간과 주파수 축의 중요한 감정적 패턴을 학습하며, 노이즈나 불필요한 정보를 효과적으로 필터링한다.

[0112] 장기 의존성 학습에서, 음성 신호에서 장기적 변화(예: 억양이나 속도의 점진적 변화)를 학습하는 데 적합하다.

[0113] Vision Transformer는 log-Mel 스펙트로그램에서 발생하는 시간적, 주파수적 변화를 효율적으로 학습할 수 있는 강력한 구조이다. 특히 Self-Attention 메커니즘을 통해 중요한 정보를 강조하고 장기 의존성을 학습함으로써, 음성 감정 인식의 성능을 크게 향상시킬 수 있다.

[0114] 셀프 어텐션 레이어는 입력 시퀀스 X로부터 query Q, key K, value V를 도출하기 위해 사용되는 세 개의 학습 가능한 가중치 행렬로 구성된다. 출력 어텐션 attention(Q,K,V)는 입력 시퀀스의 가중치이다. K^T는 K의 전치를 나타내며, d_k는 K의 차원을 나타낸다.

[0115] 도 5는 본 발명에서 감정 분류에 사용한 vision transformer의 구조를 나타낸다. Vision transformer는 6개의 헤드, 5층 깊이의 어텐션으로 이루어진 멀티헤드 어텐션으로 구성된다. 어텐션 메커니즘을 통해 분석이 끝난 특징은 MLP 헤드를 통해 감정을 나타내는 값으로 출력된다.

[0116] 도 5에 도시된 바와 같이, Vision Transformer(ViT)를 이용한 음성 감정 인식의 과정을 나타낸 것이다. 이 과정은 다음과 같이 구성된다.

[0117] 입력 데이터 (Compression Image)로, 압축된 log-Mel 스펙트로그램 이미지(128x129 크기)가 입력으로 제공된다. 이는 이전 CNN 기반 압축 단계를 통해 생성된 데이터로, 시간과 주파수 축의 중요한 정보를 포함하고 있다.

[0118] Linear Projection에서, 입력된 스펙트로그램 이미지는 Linear Projection 단계를 통해 작은 패치로 분할된다. 이 단계는 각 패치를 벡터 형식으로 변환하며, 이후 Self-Attention 메커니즘에서 처리하기 적합한 형태로 데이터가 준비된다.

[0119] Self-Attention Mechanism에서, Self-Attention 메커니즘은 입력 패치 간의 관계를 학습하고, 중요한 패치에 더 높은 가중치를 부여한다. ViT는 6개의 Attention Head와 5층 깊이의 Multi-Head Attention 구조를 사용하여, 입력 데이터의 공간적 및 시간적 특징을 효과적으로 학습한다.

[0120] MLP Head에서, Self-Attention을 통해 추출된 특징들은 MLP(Multi-Layer Perceptron) Head로 전달된다. 이 단계에서 추출된 특징은 감정 카테고리(예: "Happy," "Sad," "Angry")로 매핑된다.

[0121] Output (Emotion Classification)에서, 최종 출력은 다양한 감정 상태를 나타내는 값들로, 음성 데이터에서 추출된 감정적 특성을 나타낸다.

[0122] Self-Attention 메커니즘은 입력 데이터 내 중요한 패턴과 관계를 강조하여, 감정 인식 정확도를 높인다. 특히 음성 신호에서의 시간적 변화와 주파수 정보 간의 복잡한 상호작용을 학습하는 데 효과적이다.

[0123] 패치 기반 처리에 있어서, ViT는 입력 이미지를 작은 패치로 나누어 처리하므로, 시간과 주파수의 세부적인 특징까지 학습할 수 있다. 이는 음성 감정 인식에서 감정적 변화를 세밀하게 포착하는 데 유리하다.

[0124] 학습 효율성에 있어서, Linear Projection과 Multi-Head Attention 구조는 계산 리소스를 효율적으로 사용하면서도, 고성능 감정 인식을 가능하게 한다.

[0125] 본 발명의 모든 실험은 Windows 워크스테이션의 Windows Subsystem for Linux 2(WSL2) 도커 컨테이너 상에서 수행되었다. Windows 워크스테이션의 사양은 다음과 같다. CPU는 Intel(R) Core(TM) i7-10700k 3.8GHz, RAM은 16GB 3200Hz, GPU는 NVIDIA GeForce RTX 3070(Video RAM 8GB)이다. 개발 도구로는 Visual Studio Code와 Pytorch 2.0.1, CUDA 11.8, cuDNN 8.2.1 버전을 사용하였다.

[0126] 데이터셋으로는 Crowd Sourced Emotional Multimodal Actors(CREMA) 데이터셋을 사용하였으며, 이 데이터셋은 7,442개의 다양한 연령과 성별의 배우가 기쁨, 역겨움, 슬픔, 행복, 중립, 놀람의 6가지 감정에 대해 연기하는 영상으로 이루어져 있다. 해당 데이터셋의 7,442개의 음성 데이터를 80%:20%의 비율로 학습 세트와 테스트 세트로 분할하여 각각을 딥러닝 모델의 학습 및 성능 평가에 활용하였다.

[0127] 본 실시예에 따른 음성 감정 인식에서의 어텐션 노이즈 감소를 위한 CNN기반의 Log-Mel 스펙트로그램 이미지 압축을 위한 시스템의 실험 결과와 관련하여, 본 발명에서 제안하는 어텐션 노이즈 감소를 위한 CNN 기반의 log-Mel 스펙트로그램 이미지 압축 기법의 객관적인 성능을 평가하기 위해 표 1과 같이 음성 감정 인식 정확도를 비교하였다. 모든 딥러닝 모델은 cross entropy loss를 이용하여 학습되었다. [식 4]는 n개의 class에 대한 cross entropy loss의 계산식을 나타낸다.

[0128] [식 4]

$$L_{CE} = -\sum_{i}^{n} t_i \log(p_i)$$

[0129]

[0130] 여기서, t_i는 딥러닝 모델의 정답으로 삼는 I에 해당하는 감정 레이블, p_i는 딥러닝 모델 출력의 i번째 감정에 대한 소프트맥스 확률을 나타낸다.

[0131] [표 1]

표 1. CREMA 데이터셋에 대한 음성 감정 인식 정확도 비교 결과

Table 1. Accuracy comparison results about CREMA dataset

No	Method	Accuracy ↑
1	AST [4]	67.81%
2	SepTr [5]	70.47%
3	Ours(nearest)	56.41%
4	Ours(CNN-based)	86.83%

[0132]

[0133] 표 1의 1번 AST 기법과 2번 SepTr 기법은 압축하지 않은 이미지를 이용하여 vision transformer 구조의 네트워크를 통해 음성 감정 인식을 수행한다. 보다 객관적인 성능을 평가하기 위해 인접한 픽셀 값을 이용하는

nearest 보간을 사용하여 128x1001 크기의 이미지를 128x128 크기로 압축한 경우도 평가하여 표 1의 3번에 나타내었다. Nearest 보간으로 이미지 압축한 경우는 입력 이미지가 본 발명에서 사용한 vision transformer로 바로 입력된다. 평가 결과 제안하는 CNN 기반의 이미지 압축 기법이 86.83%의 가장 높은 정확도를 나타내었다. 기존의 정사각형 커널을 사용하는 CNN과 다르게 시간 축에 대하여 수평으로 넓은 커널을 사용하는 CNN 기반의 이미지 압축 기법이 log-Mel 스펙트로그램 분석에 더 적합하다고 사료된다.

[0134] 또한, transformer의 어텐션 메커니즘을 사용한 음성 감정 인식에서 더 높은 정확도를 달성하였다는 것은 감정 인식에 중요하지 않은 부분에 대한 잘못된 어텐션으로 인하여 생기는 어텐션 노이즈가 감소되었음을 의미한다. [표 1]은 CREMA 데이터셋에 대한 음성 감정 인식 정확도 비교 결과를 나타낸다.

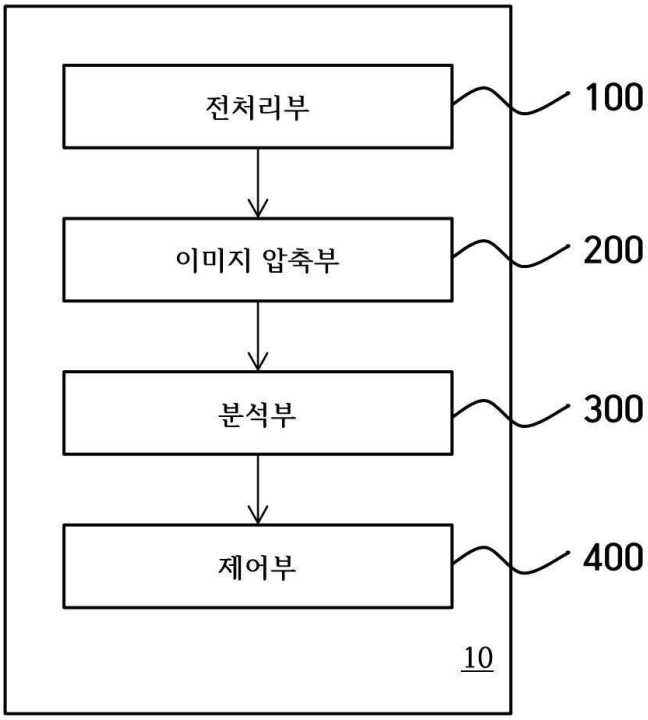
[0135] 본 발명은 CNN을 활용한 이미지 압축 방식을 통해 감정 기반 스펙트로그램 분석에서 시간적(수평적) 주파수 패턴의 변화를 효과적으로 파악할 수 있음을 보였다. 이러한 접근은 vision transformer가 어텐션 메커니즘을 수행하는 과정에서 발생할 수 있는 오류를 줄이는 동시에, 주파수 특징의 보다 정밀한 학습을 가능하게 한다는 점에서 큰 의의를 가진다. 특히, CNN의 컨볼루션 가중치는 스펙트로그램 이미지의 시간 축(x축)에 불필요한 정보가 포함되더라도 이를 효율적으로 걸러내고, 중요한 정보를 강조하는 데 기여하였다.

[0136] 이는 CNN이 지닌 로컬 필터링 특성이 감정 인식에 있어 유의미한 주파수 특징을 선택적으로 학습하도록 지원했음을 시사한다. 또한, 본 발명의 접근법은 기존의 vision transformer 모델이 단독으로 활용될 때 직면할 수 있는 어텐션 메커니즘의 비효율성을 보완할 수 있는 가능성을 제시한다. 구체적으로, CNN 기반의 시간적 이미지 압축은 vision transformer 모델이 전역적인 정보뿐만 아니라 지역적인 주파수 패턴에도 보다 민감하게 반응할 수 있도록 설계되어, 궁극적으로 감정 인식의 정확도 향상을 이끌어냈다.

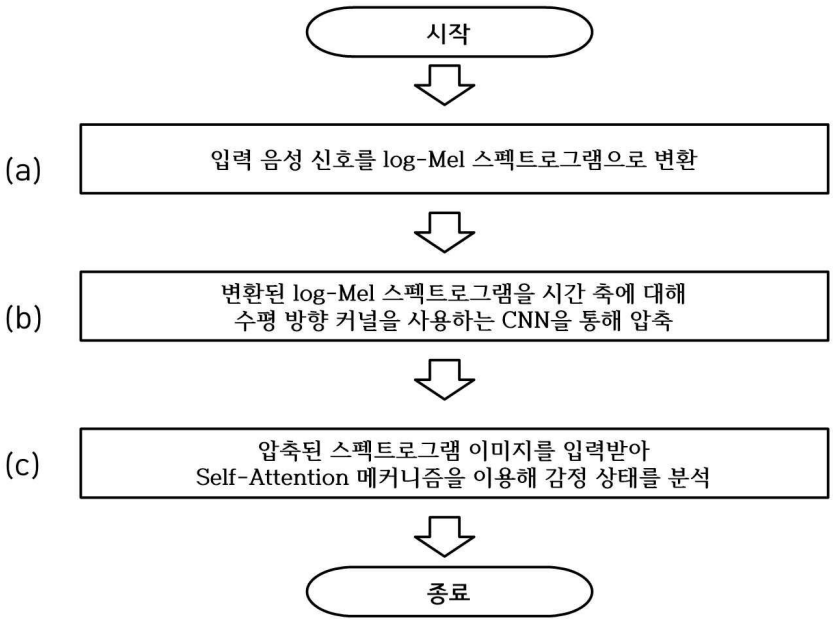
[0137] 향후, CNN과 vision transformer의 융합 구조를 확장하여 다양한 데이터 도메인에서의 일반화 가능성을 탐구할 필요가 있다. 추가적으로, 주파수 도메인 외에도 공간적 특징 및 시간적 연속성을 함께 분석함으로써 다차원적인 감정 분석 모델의 개발이 가능할 것으로 기대된다.

도면

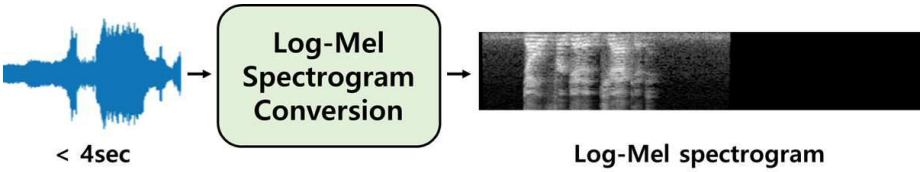
도면1



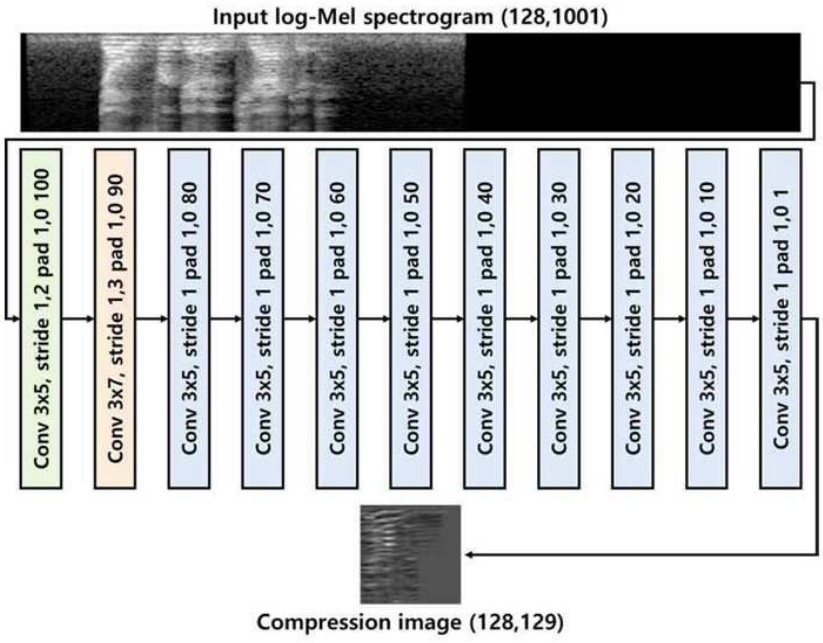
도면2


시작
(a)
입력 음성 신호를 log-Mel 스펙트로그램으로 변환
(b)
변환된 log-Mel 스펙트로그램을 시간 축에 대해
수평 방향 커널을 사용하는 CNN을 통해 압축
(c)
압축된 스펙트로그램 이미지를 입력받아
Self-Attention 메커니즘을 이용해 감정 상태를 분석
종료

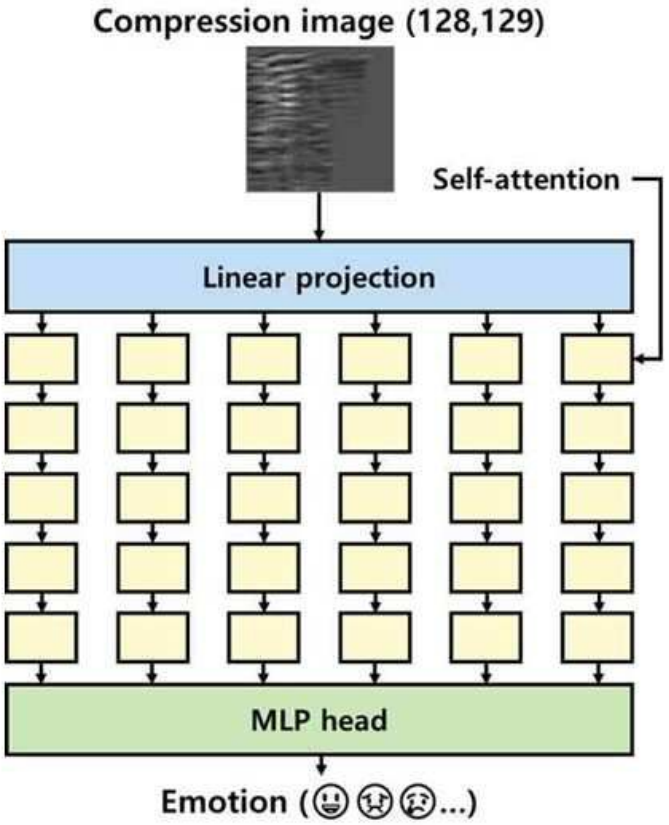
도면3


Log-Mel
Spectrogram
Conversion
< 4sec
Log-Mel spectrogram

도면4


Input log-Mel spectrogram (128,1001)
Conv 3x5, stride 1,2 pad 1,0 100
Conv 3x7, stride 1,3 pad 1,0 90
Conv 3x5, stride 1 pad 1,0 80
Conv 3x5, stride 1 pad 1,0 70
Conv 3x5, stride 1 pad 1,0 60
Conv 3x5, stride 1 pad 1,0 50
Conv 3x5, stride 1 pad 1,0 40
Conv 3x5, stride 1 pad 1,0 30
Conv 3x5, stride 1 pad 1,0 20
Conv 3x5, stride 1 pad 1,0 10
Conv 3x5, stride 1 pad 1,0 1
Compression image (128,129)

도면5


Compression image (128,129)
Self-attention
Linear projection
MLP head
Emotion (😀😭😢...)