한국음향학회지 제40권 제3호 pp. 228~233 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.3 (2021)
https://doi.org/10.7776/ASK.2021.40.3.228

pISSN : 1225-4428
eISSN : 2287-3775

3차원 합성곱 양방향 게이트 순환 신경망을 이용한 음악 템포 자극에 따른 다채널 뇌파 분류 방식

Multi-channel EEG classification method according to music tempo stimuli using 3D convolutional bidirectional gated recurrent neural network

김민수,[1] 이기용,[1] 김형국[1†]

(Min-Soo Kim,[1] Gi Yong Lee,[1] and Hyoung-Gook Kim[1†])

[1]광운대학교 전자융합공학과

(Received March 17, 2021; revised May 10, 2021; accepted May 21, 2021)

초 록: 본 논문에서는 다양한 음악 템포 자극에 따라 변화하는 다채널 ElectroEncephaloGraphy(EEG)의 특징을 추출하고 분류하는 방식을 제안한다. 제안하는 방식에서 3차원 합성곱 양방향 게이트 순환 신경망은 전처리 과정 통해 변환된 3차원 EEG 입력 표현으로부터 시공간 및 긴 시간 종속적 특징을 추출한다. 실험 결과는 제안된 템포 자극 분류 방식이 기존의 방식보다 우수하며 음악 기반 뇌-컴퓨터 인터페이스를 구축할 수 있는 가능성을 보여준다.
핵심용어: 뇌파, 템포 자극, 3차원 합성곱 양방향 게이트 순환 신경망, 게이트 순환 유닛

ABSTRACT: In this paper, we propose a method to extract and classify features of multi-channel Electro-Encephalo Graphy (EEG) that change according to various musical tempo stimuli. In the proposed method, a 3D convolutional bidirectional gated recurrent neural network extracts spatio-temporal and long time-dependent features from the 3D EEG input representation transformed through the preprocessing. The experimental results show that the proposed tempo stimuli classification method is superior to the existing method and the possibility of constructing a music-based brain-computer interface.
Keywords: ElectroEncephaloGraphy (EEG), Tempo stimuli, 3D convolutional bidirectional gated recurrent neural network, Gated recurrent unit

PACS numbers: 43.64.Lj, 43.75.Cd

I. 서 론

초고속 네트워크 환경의 발전과 함께 스마트폰 기반의 라이프 스타일 변화로 디지털화된 음원을 온라인에서 실시간으로 들을 수 있는 음악 스트리밍 서비스가 확산되었다. 이에 따라 디지털 음원의 정보를 분석하여 사용자가 원하는 음악의 정보를 찾아주는 음악 정보 검색(Music Information Retrieval, MIR) 시스템에 대한 연구가 활발히 진행되고 있다.

MIR 분야에서 주목 받는 연구 주제 중 하나인 음악 분류에는 제목, 가수, 장르와 같은 노래의 다양한 요소가 적용되었다. 최근에는 사용자 중심의 MIR 시스템을 구축하기 위해 음악에 대한 인간의 인지 메커니즘을 모델링하는 연구가 도입되고 있다.[1] 또한 뇌 반응에 대한 관심이 높아지면서 뇌 활동의 가장 두드러진 변화를 일으키는 음악적 특성을 파악하려는 연구

†**Corresponding author:** Hyoung-Gook Kim (hkim@kw.ac.kr)
Department of Electronics Convergence Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Republic of Korea
(Tel: 82-2-940-5574, Fax: 82-2-913-5006)

가 증가하고 있다.[2] 이러한 특성에는 박자, 분위기, 템포 및 악기의 유무가 있으며 그중에서 템포는 곡의 빠르기를 지칭하는 용어로 주의력, 시간 지각, 의사 결정과 같은 인간의 인지 능력에 영향을 미칠 수 있다.[3] 이를 기반으로 음악의 템포 변화에 따라 반응하는 ElectroEncephaloGraphy(EEG) 기록으로부터 인간이 인지한 템포 자극을 해석할 수 있다면 MIR 시스템에 적용할 수 있다. 그러나 EEG를 이용한 MIR 시스템의 경우 아직 미개척 영역이며 대부분의 연구는 심층 신경망을 기반으로 한 뇌파의 감정 인식에 초점을 맞추고 있다.[4] 그중 Salama *et al.*[5] 는 3D Convolutional Neural Network(3D CNN)을 사용하여 다채널 EEG의 채널 간 상관관계를 학습하는 방식을 제안하였다. 3D CNN은 인접한 시간 영역의 특징 정보를 학습할 수 있지만 긴 시퀀스에서 시간에 따른 특징 정보의 상관관계를 고려할 수 없었다. 반면 순환 신경망의 한 종류인 Gated Recurrent Unit(GRU)[6]는 Long Short Term Memory(LSTM)보다 적은 매개변수를 통해 빠른 학습 속도를 보이며 순환 신경망의 고질적인 문제인 장기 의존성 문제를 해결하여 우수한 성능을 보였다.

이에 본 논문에서는 음악을 인지하는 동안 기록된 다채널 EEG로부터 시공간 특징과 시간 종속적 특성을 모두 학습하기 위해 3차원 합성곱 양방향 게이트 순환 신경망(3Dimensional Convolutional Bidirectional Gated Recurrent Neural Network, 3D CBGRNN)을 이용한 템포 자극 분류 방식을 제안한다.

본 논문의 구성은 다음과 같다. II장에서는 제안하는 템포 자극 분류 방식의 구조에 대해 설명한다. III장에서는 실험 결과를 제시하고 마지막으로 IV장에서는 결론을 맺는다.

II. 템포 자극 분류 방식의 구조

Fig. 1은 본 논문에서 제안하는 템포 자극 분류 방식의 기본적인 프레임워크를 보여준다.

우선 피험자는 다채널 EEG 측정 장치를 착용한 상태에서 다양한 Beats Per Minute(BPM)을 가진 노래를 들으며 EEG를 측정한다. 이후 125 채널 EEG 전극 캡을 통해 수집된 모든 채널의 EEG 신호를 30 s 길이로 분할한다. 데이터베이스에 저장된 각 EEG 데이터는

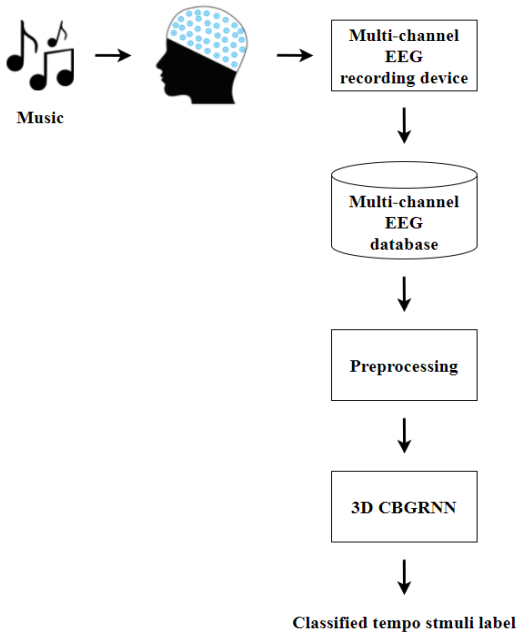

Fig. 1. (Color available online) Simplified framework for tempo stimuli classification from EEG.

전처리 과정을 거쳐 3차원 구조로 변환되어 3D CNN과 BGRNN으로 구성된 3D CBGRNN에 입력된다. 3D CBGRNN 모델은 다채널 EEG의 시공간 및 시간 종속적 특징을 추출하고 이를 통해 템포 자극 레이블 중 하나로 분류한다.

2.1 전처리 과정

시간 정보와 채널 별 위치 정보를 모두 포함하는 다채널 EEG 신호에서 템포 자극에 해당하는 특징을 추출하기 위해서 3차원 구조의 입력 표현으로 변환하는 전처리 과정이 필요하다. 이를 위해, Fig. 2와 같이 3D EEG 입력 표현으로 변환하는 과정을 제시한다. 모든 채널에 해당하는 30 s 길이의 1차원 EEG 신호를 1 s 길이의 윈도우를 사용하여 프레임 단위로 분할한다. 이때 한 채널의 프레임의 수는 30개이며 각 프레임에는 125개의 샘플이 존재한다. 전체 125 채널의 모든 프레임은 순차적으로 첫 번째 공간 영역의 축을 따라 추가되어 높이는 프레임 수이고 너비는 프레임의 샘플 수인 2차원 배열을 형성한다. 이처럼 채널 × 프레임 × 샘플 구조로 형성된 각 3D EEG 입력 표현은 시공간 특징을 학습하기 위해 3D CNN에 입력된다.

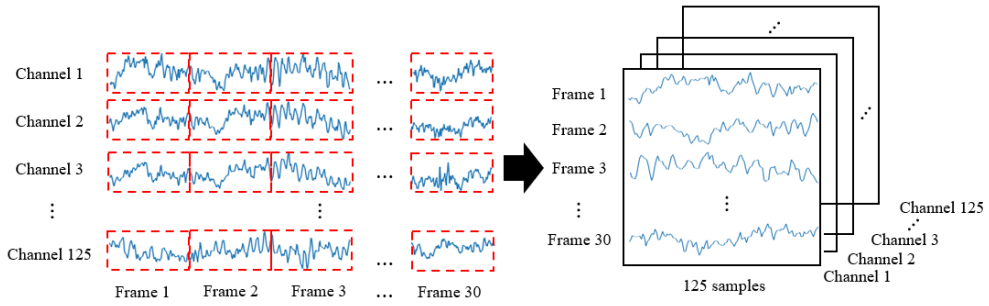

Fig. 2. (Color available online) 3D Input representation of EEG signal.

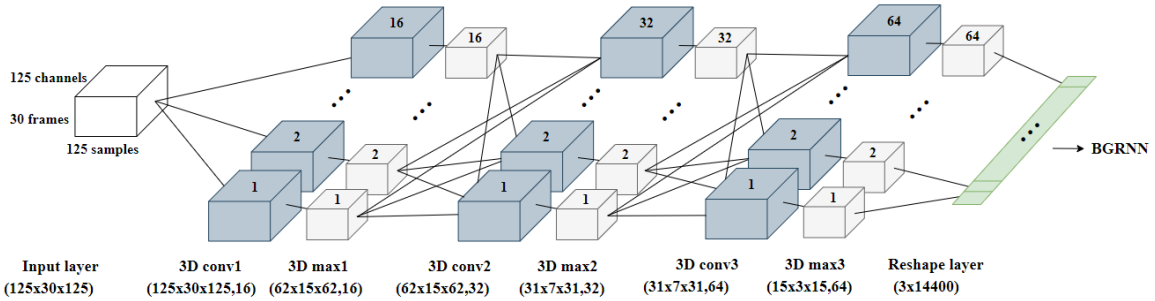

Fig. 3. (Color available online) Architecture of the proposed 3D CNN.

2.2 3차원 합성곱 신경망

3D CNN은 기존 CNN에 합성곱 및 Pooling 연산을 변형한 구조로 비디오와 같은 긴 시퀀스의 시공간 특징을 학습하기 위해 사용된다. 이를 반영하여 시공간 및 시간 종속적 특징을 효과적으로 모델링하기 위해 다채널 EEG 신호에 3D CNN을 적용한다. 3D 합성곱 필터는 입력된 3D EEG 표현으로부터 합성곱 연산을 통해 인접한 시공간 영역의 상관관계를 갖는 3D 특징 표현을 추출한다.

Fig. 3은 본 논문에서 제안된 3D CNN의 구조를 보여준다. 3D CNN은 연속으로 세 번 연결된 3D 합성곱 층과 3D Max-Pooling 층 그리고 reshape 층으로 구성된다. 각 3D 합성곱 층은 3D Convolutional filter(Conv), Batch Normalization(BN), Rectified Linear Units(ReLU) 활성화 함수로 구성된다. 첫 번째 3D 합성곱 층에서, 3D 합성곱 필터는 입력된 3D EEG 표현에 적용되어 합성곱 연산을 수행한다. 3D 합성곱 연산은 다음과 같은 식으로 표현된다.

$$R(i,j,k) = \sum_x \sum_y \sum_z C(x,y,z) * V(i-x,j-y,k-z), \quad (1)$$

여기서 R은 3D 합성곱 연산의 출력이고, C는 $x \times y \times z$ 크기의 3D 합성곱 필터이다. 또한 V는 3D 합성곱 필터와 연산되는 입력 3D EEG 표현이며 필터 C의 크기는 항상 V의 크기보다 작게 설정해야 한다.

3D 합성곱 필터 C는 입력 3D 표현 V에 포함된 모든 3D 영역과의 합성곱 연산을 통해 각 영역의 상관관계와 함께 3D 특징 표현 R을 출력한다. 이때 합성곱 필터 C의 수만큼 다양한 3D 특징 표현을 추출할 수 있다. 추출된 모든 3D 특징 표현은 BN과 ReLU 활성화 함수가 적용된 후 3D Max-Pooling 층에 입력되어 3D Pooling 영역 중에 가장 뚜렷한 시공간 특징을 추출한다. 동일한 과정을 반복함으로써, 연속된 시공간 영역에서 뚜렷한 상관관계만을 포함하는 3D 시공간 특징을 추출할 수 있다.

제안된 3D CNN의 모든 3D 합성곱 필터와 3D Max-Pooling은 각각 $3 \times 3 \times 3$과 $2 \times 2 \times 2$의 크기를 가지며, 3D EEG 입력의 시공간 특징을 추출하기 위해 사용된다. 또한 3D 합성곱 필터의 수를 16, 32, 64로 설정하여 채널 간 다양한 상관관계를 효율적으로 학습할 수 있도록 한다. 세 번째 3D Max-Pooling 층에서 추출된 64개의 3D 시공간 특징맵은 프레임 축 정보를 유지하고 인접한 시간 및 공간 정보에서 추출된

뚜렷한 시공간 특징을 포함한다. 64개의 3D 시공간 특징은 reshape 층을 통해 순차적으로 연결되고 크기가 3 × 14400인 특징 표현의 형태로 출력된다.

2.3 양방향 게이트 순환 신경망

3D CNN을 통해 출력된 3D 특징 표현은 인접한 시공간 영역의 상관관계를 포함하고 있지만 긴 시간 동안의 시간 특징 정보를 고려할 수 없다. 따라서 3D CNN에서 출력된 3(프레임) × 14400(특징 벡터) 크기의 특징 표현이 BGRNN에 입력되어 시계열 시퀀스로부터 과거 및 미래 정보의 상관관계를 최대한 활용함으로써 더 긴 시간 종속성을 효과적으로 모델링한다.

본 논문에서 사용된 BGRNN은 Fig. 4와 같은 구조로 순방향 시퀀스와 역방향 시퀀스로 나누어진 2개의 GRU로 구성된다. 이를 통해, BGRNN은 과거와 미래 정보를 최대한 활용하여 현재 상태 특징 정보를 학습할 수 있다. 또한, 순방향 및 역방향 GRU 은닉층의 셀 크기는 128로 설정함으로써 각 시퀀스에서 256개의 더 긴 시간 특징이 추출되며, 이는 각각 14400개의 특징 벡터를 포함한다.

BGRNN의 GRU는 LSTM을 변형한 구조이며 Fig. 5와 같이 GRU의 구조는 각각 메모리 내부 및 외부로 유입되는 정보를 제어하는 업데이트 게이트와 리셋 게이트로 구성된다. GRU 은닉층의 연산 과정은 다음과 같은 계산 과정을 갖는다.

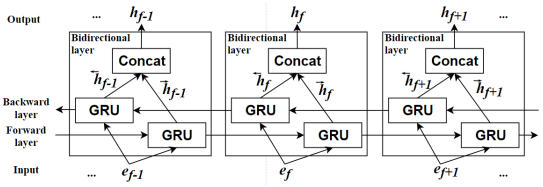

Fig. 4. Bidirectional GRU network structure.

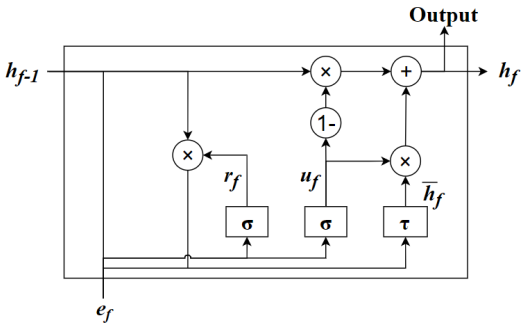

Fig. 5. Internal computing structure of GRU.

$$r_f = \sigma(W_r h_{f-1} + U_r e_f), \tag{2}$$

$$u_f = \sigma(W_u h_{f-1} + U_u e_f), \tag{3}$$

$$\bar{h}_f = \tau(W h_{f-1} \cdot r_f + U e_f), \tag{4}$$

$$h_f = (1-u_f) \cdot h_{f-1} + u_f \cdot \bar{h}_f, \tag{5}$$

여기서 e_f는 f번째 프레임의 특징 벡터 e, r과 u은 각각 리셋 게이트와 업데이트 게이트를 나타낸다. 또한 h, W, U, σ, τ는 각각 은닉층의 출력, 은닉층과 은닉층 사이의 가중치, 입력층과 은닉층 사이의 가중치, 시그모이드 함수, 하이퍼볼릭 탄젠트 함수를 나타낸다. 리셋 게이트는 Eq. (2)에 표시된 것처럼 시그모이드 함수를 통해 0과 1 사이의 출력 값을 가지며, Eq. (4)에서 현재 상태의 입력과 이전 메모리의 정보를 제어하기 위해 사용된다. Eq. (3)에 표시된 업데이트 게이트 출력 값도 0과 1 사이이며, 이전 메모리에 대한 현재 입력의 비율을 결정하기 위해 Eq. (5)의 은닉층 출력 계산 과정에 적용된다. 따라서 GRU는 이전 상태의 정보를 효과적으로 사용하여 현재 상태의 특징 정보를 결정할 수 있으며 역방향 시퀀스를 통해 다음 상태의 정보까지 고려하여 강력한 시간적 상관관계를 반영한다.

다음으로 BGRNN의 은닉층으로부터 출력된 주석 벡터는 Softmax 층에 입력되어 각 템포 자극 클래스의 확률을 계산하고 가장 높은 확률에 해당하는 템포 자극 레이블을 예측 결과로 출력한다. 모델의 학습을 위해 아담 최적화 방식을 통해 가중치를 갱신하고 예측된 클래스와 실제 클래스의 교차 엔트로피 오류를 최소화하였다.

III. 실 험

3.1 실험 데이터

본 논문의 실험에서는 뇌파 반응 및 행동 반응을 포함하는 공개 데이터 세트인 Naturalistic Music EEG Dataset-Tempo(NMED)[2]를 사용하여 제안된 템포 자

극 분류 방식을 평가하였다. 청각 자극으로 사용된 상용 음악 10곡은 모두 270 s ~ 300 s 길이이며 서양 전통 음악의 다양한 장르와 템포를 포함한다. EEG 신호는 20명의 피험자로부터 125 Hz 샘플링 속도로 125 채널 전극 캡에 의해 기록되었다. 또한, 행동 반응 데이터는 EEG 측정을 마친 피험자가 1과 9 사이의 행동 등급으로 각 노래에 대한 친숙도와 즐거움의 수준을 평가한 기록과 피험자가 노래의 짧은 발췌를 다시 듣고 응답한 탭핑 기록으로 구성된다.

NMED-T의 템포 자극을 적절히 반영하는 EEG 신호만 선택하여 템포 자극을 분류하는 기본 실험을 수행하기 위해 다음과 같은 기준을 적용했다. 1) 뇌 활동을 자극하여 EEG 기록에 반영되는 익숙한 노래를 피하기 위해, 친숙도 등급이 4 미만인 피험자의 EEG 데이터만 선별했다. 2) 각 곡마다 평가된 즐거움 등급의 차이가 뚜렷하여 음악 템포 자극을 명확하게 인지한 피험자의 EEG 데이터만을 선별하였다. 3) 탭핑 데이터를 Hz 단위로 변환하여 음원의 템포 주파수(Hz)와 비교하여 오류가 적은 EEG 데이터만 선별하였다. 4) 피험자의 피로도로 인해 발생할 수 있는 오류를 방지하기 위해 즐거움 등급이 각 곡에 대한 평균적인 수치보다 편차가 매우 큰 경우를 제외한 EEG 데이터만이 선별되었다.

템포 자극이 높은 순서로 EEG 데이터베이스를 구축하기 위해서 데이터 선별이 완료된 후 20대 남성 3명, 여성 2명으로 구성된 새로운 평가자 5명에게 노래 10곡을 듣고 템포 자극의 정도를 1 ~ 5 사이의 점수로 기록하도록 요청하였다. 평균 템포 자극 점수가 높은 순서대로 피험자 10명에 대한 4, 6, 8곡에 해당하는 3개의 EEG 데이터베이스가 구축되었다. 이때 피험자 1명이 노래 1곡을 들으면서 측정한 270 s 길이의 EEG 신호를 30 s 길이로 분할한 후 9개의 EEG 데이터를 확보하였고 각 곡에 대한 피험자 10명의 EEG 데이터는 총 90개로 구성된다. 다음으로 뇌 반응을 유도하는 6가지 템포 자극 클래스를 설정하여 각 EEG 데이터에 해당하는 템포 자극 레이블을 색인하였다. 4곡으로 구성된 데이터베이스에는 4개의 템포 자극에 따라 기록된 360개의 EEG 데이터가, 6곡으로 구성된 데이터베이스에는 6개의 템포 자극에 따라 기록된 540개의 EEG 데이터가 포함된다. 8곡으로 구성된 데이터베이스에는 동일한 템포 자극 클래스에 속하는 노래가 존재하며 6개의 템포 자극에 의해 기록된 720개의 EEG 데이터를 포함한다.

3.2 측정 방식

본 논문에서는 데이터의 편중을 방지하기 위해 4-fold 교차 검증 방식을 이용하였으며 제안하는 3D CBGRNN 방식을 검증하기 위해 다른 신경망 방식의 결과와 성능을 비교하였다. 먼저 2D CNN은 3개의 2차원 합성곱 층과 2D Max-Pooling 층으로 구성된다. 모든 2차원 Conv와 2D Max-Pooling의 크기는 각각 3×3, 2×2이며 각 합성곱 필터의 수는 16, 32, 64개로 설정하였다. 다음으로 3D CNN은 앞서 설명한 3D CNN의 구조를 동일하게 적용하여 실험을 진행하였다. ResNet의 경우 잔류 학습 방식을 적용하여 입력과 출력의 차이인 잔차를 학습하는 신경망으로 4개의 잔류 블록을 사용하여 실험을 진행하였다. 또한 본 실험에서 BGRNN은 128개의 셀 개수를 적용하였고 2D CBGRNN은 2D CNN과 BGRU를 결합하여 실험을 진행하였다. 모든 신경망은 learning rate는 0.001, batch size는 32로 설정하여 100 Epoch까지 학습을 진행했다.

3.3 실험 결과

Table 1은 본 논문에서 진행한 실험의 결과를 보여준다. 본 논문의 실험에서 제안된 3D CBGRNN 방식은 4, 6, 8곡에 대한 3개의 EEG 데이터 세트와 NMED-T의 모든 노래의 데이터베이스에 대해 가장 우수한 결과를 보여주었다. 그중에서 4곡에 대한 EEG의 템포 자극 분류에서 86.7 %로 가장 높은 정확도를 보였다. 6곡에 대한 EEG 템포 자극 분류 정확도는 81.3 %로 4곡에 대한 분류 결과보다 5.4 % 낮았고, 8곡에 대

Table 1. Result of tempo stimuli classification.

Method	Accuracy			
	4 song	6 song	8song	all
2D CNN	80.2 %	67.3 %	59.9 %	47.8 %
3D CNN	81.3 %	68.9 %	61.7 %	49.6 %
ResNet	79.5 %	67.2 %	59.7 %	47.5 %
BGRNN	82.9 %	70.6 %	63.2 %	50.6 %
2D CBGRNN	84.8 %	74.7 %	66.8 %	52.4 %
3D CBGRNN	86.7 %	81.3 %	71.7 %	54.8 %

한 EEG 템포 자극 분류 정확도인 71.7 %보다는 9.6 % 높았다. 가장 낮은 분류 정확도는 54.8 %로 NMED-T의 모든 곡에 대한 EEG의 템포 자극 분류에서 나타났다. 3D CNN 방식은 전반적으로 2D CNN 방식보다 낮은 분류 결과를 보였다. 이러한 결과를 통해 채널 간 위치의 상관관계를 고려할 수 있는 3D CNN이 2D CNN보다 다채널 EEG의 템포 자극 분류에 효과적임을 확인할 수 있다. 또한 2D CBGRNN 방식의 경우 BGRNN 방식보다 우수한 성능을 보였지만 3D CBGRNN에 비해 낮은 성능을 보였다. 이를 통해 또한 제안하는 방식이 압축된 시공간 특징을 추출하여 BGRU의 기울기 손실 문제와 장기 의존성 문제를 해결함으로써 성능을 더욱 향상시켰음을 알 수 있다. 가장 낮은 결과는 모두 ResNet 방식에서 나타났다.

IV. 결 론

본 논문에서는 3차원 합성곱 양방향 게이트 순환 신경망을 이용하여 EEG의 템포 자극을 분류하는 방식을 제안하였다. 실험 결과는 제안된 방식이 다채널 EEG의 시공간 특징뿐만 아니라 시간 종속적 특징까지 효과적으로 모델링하여 템포 자극 분류의 성능을 향상시켰음을 보여준다. 향후 본 연구를 바탕으로 템포 자극 클래스와 노래의 곡 수를 늘려 데이터베이스의 규모를 확장하고, 최근 시퀀스 데이터의 문맥 정보를 학습하기 위해 사용된 주의집중 메커니즘을 적용하여 성능을 향상시켜 음악 템포 자극을 적용한 뇌-컴퓨터 인터페이스를 구현하고자 한다.

감사의 글

본 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원과 2021년도 광운대학교 우수연구자 지원 사업의 지원을 받아 수행된 기초연구사업임(NRF-2018R1D1A1B07041783).

References

1. T. Greer, B. Ma, M. Sachs, A. Habibi, and S. S. Narayanan, “A multimodal view into music's effect on human neural, physiological, and emotional experience,” Proc. ACM Int. Conf. Multimedia, 167-175 (2019).
2. S. Losorelli, D. T. Nguyen, J. P. Dmochowski, and B. Kaneshiro, “NMED-T: A tempo-focused dataset of cortical and behavioral responses to naturalistic music,” Proc. ISMIR. 339-346 (2017).
3. I. N. Clark and J. Tamplin, “How music can influence the body: Perspectives from current research,” Voices: A World Forum for Music Therapy, **16**, 1-15 (2016).
4. J. L. Hsu, Y. L. Zhen, T. C. Lin, and Y. S. Chiu, “Affective content analysis of music emotion through EEG,” Multimed. Syst. **24**, 195-210 (2018).
5. E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Shalaby, “EEG-based emotion recognition using 3D convolutional neural networks,” Int. J. Adv. Comput. Sci. Appl. **9**, 329-337 (2018).
6. Y. Guo, W. Liu, D. wei, and Q. Chen, “Emotional recognition based on EEG signals comparing long-term and short-term memory with gated recurrent unit using Batch Normalization,” Proc. Int. Conf. MEET. 1-9 (2019).

저자 약력

▸ 김 민 수 (Min-Soo Kim)

2021년 2월 : 광운대학교 전자융합공학과 학사
2021년 3월 ~ 현재 : 광운대학교 전자융합공학과 석사과정

▸ 이 기 용 (Gi Yong Lee)

2020년 2월 : 광운대학교 전자융합공학과 학사
2020년 3월 ~ 현재 : 광운대학교 전자융합공학과 석사과정

▸ 김 형 국 (Hyoung-Gook Kim)

1999년 ~ 2002년 : 독일 SIEMENS/Cortologic AG 책임연구원
2002년 ~ 2005년 : 독일 베를린 공과대학교 Assistant Professor
2005년 ~ 2007년 : 삼성종합기술원 수석연구원
2007년 3월 ~ 현재 : 광운대학교 전자융합공학과 교수