J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

청크 기반 시계열 음성의 감정 인식 연구☆

A Study on Emotion Recognition of Chunk-Based Time Series Speech

신 현 삼[1] 홍 준 기[2*] 홍 성 찬[3]
Hyun-Sam Shin Jun-Ki Hong Sung-Chan Hong

요 약

최근 음성 감정 인식(Speech Emotion Recognition, SER)분야는 음성 특징과 모델링을 활용하여 인식률을 개선하기 위한 많은 연구가 진행되고 있다. 기존 음성 감정 인식의 정확도를 높이기 위한 모델링 연구 이외에도 음성 특징을 다양한 방법으로 활용하는 연구들이 진행되고 있다. 본 논문에서는 음성 감정이 시간 흐름과 연관이 있음을 착안하여 시계열 방식으로 음성파일을 시간 구간별로 분리한다. 파일 분리 이후, 음성 특징인 Mel, Chroma, zero-crossing rate (ZCR), root mean square (RMS), mel-frequency cepastral coefficients (MFCC)를 추출하여서 순차적 데이터 처리에 사용하는 순환형 신경망 모델에 적용하여 음성 데이터에서 감정을 분류하는 모델을 제안한다. 제안한 모델은 librosa를 사용하여 음성 특징들을 모든 파일에서 추출하여, 신경망 모델에 적용하였다. 시뮬레이션은 영어 데이터 셋인 Interactive Emotional Dyadic Motion Capture (IEMOCAP)을 이용하여 recurrent neural network (RNN), long short-term memory (LSTM) and gated recurrent unit(GRU)의 모델들의 성능을 비교 및 분석하였다.

☞ 주제어 : RNN, GRU, LSTM, 음성 감정 인식, 음성 특징

ABSTRACT

Recently, in the field of Speech Emotion Recognition (SER), many studies have been conducted to improve accuracy using voice features and modeling. In addition to modeling studies to improve the accuracy of existing voice emotion recognition, various studies using voice features are being conducted. This paper, voice files are separated by time interval in a time series method, focusing on the fact that voice emotions are related to time flow. After voice file separation, we propose a model for classifying emotions of speech data by extracting speech features Mel, Chroma, zero-crossing rate (ZCR), root mean square (RMS), and mel-frequency cepstrum coefficients (MFCC) and applying them to a recurrent neural network model used for sequential data processing. As proposed method, voice features were extracted from all files using 'librosa' library and applied to neural network models. The experimental method compared and analyzed the performance of models of recurrent neural network (RNN), long short-term memory (LSTM) and gated recurrent unit (GRU) using the Interactive emotional dyadic motion capture Interactive Emotional Dyadic Motion Capture (IEMOCAP) english dataset.

☞ keyword : RNN, GRU, LSTM, Speech Emotion Recognition, Voice Feature

1. 서 론

인간의 음성을 이해하고 감정을 인식하는 기술은 4차 산업의 핵심인 인공지능의 가장 관심 영역이다. 최근 음성 감정 인식 분야는 신경망 모델별 학습과 입력되는 음성의 특징을 다양하게 추출하여 성능을 측정하는 연구 성과를 내고 있다. 또한, 최근 음성 특성을 다양화하여 높은 정확도를 보여주는 연구가 많이 진행되고 있다. Artificial Neural Network(ANN), CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory)과 같은 딥러닝 기술을 사용하여 이미지[1-3], 음성인식[4], 자연어 처리 등 많은 분야에서 인공지능 모델을 사용한 연구가 활발히 진행되고 있다 [5]. CNN과 RNN으로 시간 영역의 음성 신호를 감정의 연속 순환 모델에 맵핑(mapping) 시키는 end-to-end 음성 감정 인식 방법에 Attention Mechanism을 추가하여 인식률을 향상하는 모델을 제안하였다 [6]. 또한 Interactive Emotional Dyadic Motion Capture (IEMOCAP)을 사용하여

1 Department of Information and Communication, Hanshin University, Osan-si, 18101, Korea
2 Department of Smart Information Technology Engineering, Kongju National University, Cheonan-si, 31080, Korea.
3 Department of Software Convergence, Hanshin University, Osan-si, 18101, Korea
* Corresponding author (jkhong@kongju.ac.kr)
[Received 20 January 2023, Reviewed 3 February 2013(R2 6 March 2023), Accepted 8 March 2023]
☆ 본 논문은 2022년도 한국인터넷정보학회 추계학술대회 우수논문 추천에 따라 확장 및 수정된 논문임.

LSTM, RNN, SVM(Support Vector Machines), Pooling, 가중치 Pooling 등 다양한 구조를 적용하여 감정 인식 성능을 비교 분석하였다 [7]. Zhou의 연구에서는 시간 영역 프레임으로부터 SAE(Stacked Auto Encoder) 또는 DBN (Deep Belief Network) 모델을 학습하고 sigmoid로 분류하였다 [8]. 베를린 감정 음성 데이터베이스 (EMO-DB)를 사용하여 정확도를 실험하였다 [9]. 본 논문에서는 음성이 시계열로 분석될 때 의미를 가진다는 점, 음성 특성을 추출하여 딥러닝에 활용하고, 청크 방식을 이용해 성능을 확인하는 새로운 연구 방식을 제시한다. 시계열 음성 데이터를 *N*개의 청크 신호로 분리하여 SER (Speech Emotion Recognition) 신경망 모델에 입력하고 감정을 추출하는 모델을 제안하고 성능을 평가한다.

2. 제안한 청크 기반 음성 특징 추출 모델

제안된 방법은 화자의 원본 음성 파일을 가변적인 초 단위로 음성을 잘라서 각각의 음성 파일에 대해 오디오 특성을 추출해서 수집한 이 특징들을 시계열 방식으로 전체 음성 파일을 시간별로 구분하여 (이하 청크, chunk) 하여 다양한 순환 신경망 모델의 입력 값으로 사용한다. 시계열 데이터의 특징은 시간에 관해 순서가 있다는 점과 연속한 관측치는 서로 상관관계에 있다고 볼 수 있다. 다시 말해 일정한 시간 동안 수집된 일련의 의미를 가지고 있는 순차적으로 정해진 데이터 셋의 집합이라고 할 수 있다. 이러한 시계열 데이터의 목적인 시계열 법칙성을 찾고 모형화하는 것이다[12-13]. 본 논문에서는 음성의 시계열성을 활용해서 각 구간의 감정을 추출하여 전체 음성의 특징을 판단한다. 제안한 청크 기반 감정 인식 추출 모델 (Chunk-based Voice Feature Extraction, CVFE)은 시계열 오디오 데이터의 모든 청크로부터 딥러닝을 이용하여 감정을 예측한다.

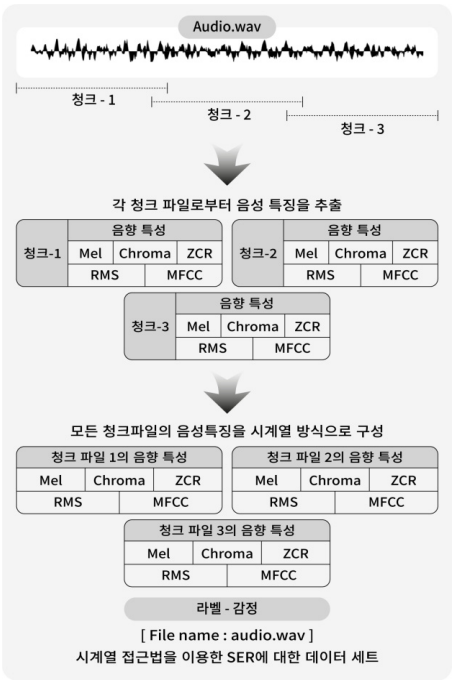
2.1 시계열 기반의 음성 특징 추출 메카니즘

제안한 CVFE 모델에서 전체 음성의 구간을 의미하는 '청크'의 길이는 2초, 3초, 5초와 같이 가변적으로 설정할 수 있으며 정해진 청크의 길이를 사용하여 전체 음성을 구별한다 [14]. 제안한 CVFE모델은 3초를 기준으로 청크를 구성하였으며 음성 데이터로부터 음성의 특징을 추출하기 위한 단계는 아래 표 1과 같다.

(표 1) 제안한 CVFE의 음성 데이터 처리 단계
(Table 1) Processing of voice data of the proposed CVFE

단 계	설 명
1단계	전체 오디오 파일의 음성 시간을 구한 다음 각각 3초씩 *n*개의 청크 파일로 분할한다. (마지막 파일은 3초에 맞게 무음으로 조정한다. 청크의 길이와 개수는 변경될 수 있다.)
2단계	각 청크 파일에서 해당 파일에 대한 음향 특징을 추출한다.
3단계	청크-1부터 청크-*n*까지 음향 특징을 모아서 시간순으로 정렬한다.

각 청크로부터 음성의 특징을 추출을 하기 위해 파이썬의 음성 분석 라이브러리인 'librosa'를 사용한다. librosa는 음악과 오디오 분석을 위한 파이썬 패키지이다. 음악 및 소리의 정보를 찾거나 분석하는 데 필요한 구성 요소를 제공하는 패키지이다 [15]. 그림 1과 같이 분할된 음성 파일을 librosa를 사용하여 음성의 특징을 추출하고 해당 음성 특징들을 별도로 분류하고 저장한다.



(그림 1) 제안한 CVFE 모델의 순서도
(Figure 1) Flowchart of proposed CVFE model

그림 1은 제안한 CVFE모델의 순서도이며 원음에서 음성파일들을 잘라내고, 잘라낸 각 청크들로부터 librosa를 사용하여 추출한 다섯 개의 음성의 특징을 추출하는 기술들은 다음 절에서 설명한다.

2.1.1 Mel Spectrogram

Mel Spectrogram은 시간에 따라 달라지는 오디오의 주파수 특성을 분석하기 위한 특징 추출기법이다. 소리의 음고 (pitch)에서 발견한 사람의 음을 인지하는 기준(threshold)을 반영한 스케일 (scale) 변환 함수이다. 아래의 식1은 기본 오디오 주파수와 mel scale 주파수의 상관 관계식으로 $mel(f)$는 mel scale로 변환된 주파수, f는 Hz 단위의 비선형적 방식의 주파수를 의미한다.

$$mel(f) = 2595 \times log(1 + \frac{f}{700}) \quad (1)$$

2.1.2 Chroma

크로마 (Chroma)는 인간 청각이 옥타브 차이가 나는 주파수를 가진 두 음을 유사음으로 인지하여 모든 스펙트럼을 12개의 옥타브로 표현하여 구분한 음성 특성이다. 즉 한 옥타브 차이가 비슷한 경우에는 색상이 비슷하며, 차이가 클수록 색상 차이가 크게 난다.

2.1.3 ZCR

ZCR (Zero Crossing Rate)은 시간 도메인의 시그널이 프레임 구간에서 부호가 바뀐 비율을 나타낸다. 서로 다른 오디오 신호를 특징짓는 데 유용하며 특히 음성 신호와 음악 신호를 분류하는 알고리즘에 많이 사용되고 있다. ZCR은 프레임 내에서 신호의 부호가 변하는 횟수를 의미하므로 일반적으로 고주파는 높은 ZCR을 보이며, 저주파는 낮은 ZCR 값을 갖는다. 임의의 프레임의 ZCR은 다음과 같이 나타낼 수 있다.

$$Z_{(i)} = \frac{1}{2\omega_L} \sum_{n=1}^{\omega_L} |sgn(x_i(n) - sgn(x_i(n-1))| \quad (2)$$

여기서 w_L은 프레임의 길이이고 i는 각 프레임의 번호다. 최종적으로 $ZCR = [Z_{(1)}, Z_{(2)}, Z_{(3)}, Z_{(4)}, Z_{(5)}]$의 형태가 된다. 함수 $sgn(x)$는 x가 0과 양수이면 1을 반환하고 x가 음수이면 -1을 반환하는 함수다.

$$sgn(x_i(n)) = \begin{cases} 1 & (x \geq 0) \\ -1 & (x < 0) \end{cases} \quad (3)$$

2.1.4 RMS

RMS (Root Mean Square)는 음성 진폭 값의 평균 제곱근이다. 본 논문에선 분할된 각 음성 프레임의 에너지를 계산한 값으로 정의한다.

$$\text{RMS} = \sqrt{\frac{\sum_{i=1}^{N} y(i)^2}{N}} \quad (4)$$

여기서 N은 프레임의 길이, y(i)는 i에서 음성 신호의 진폭 값을 의미하며 i는 각 프레임의 인덱스를 나타낸다.

2.1.5 MFCC

MFCC (Mel-Frequency Cepstral Coefficient)는 오디오 신호에서 특징을 추출하는 기법으로, 입력된 오디오 전체가 아니라 일정 구간으로 나누어 이 구간에 대한 스펙트럼을 분석하여 특징을 추출하는 기법이다. MFCC는 아래 식 5와 같이 나타낼 수 있다.

$$C_n = \sum_{n=1}^{k} (log \widehat{S}_k) cos[n(k - \frac{1}{2}) \frac{n}{k}] \quad (5)$$

여기서 $\widehat{S}_k$는 mel scale의 과정을 거쳐 나온 벡터 값, k는 필터뱅크의 index, n은 MFCC 계수의 차수를 나타낸다.

2.2 순환 신경망 모델

제안한 CVFE 모델은 시간의 변화에 따라 순서가 있는 음성 데이터를 처리하기 때문에 입출력을 시퀀스 단위로 처리하는 RNN을 사용하였다.

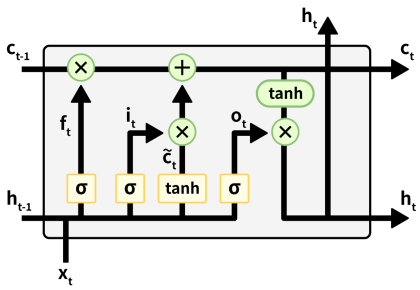
2.2.1 Recurrent Neural Network (RNN)

RNN은 입력과 출력을 시퀀스 단위로 처리하는 모델로 히든 노드가 방향을 가진 엣지로 연결되는 순환구조를 이루는 신경망이다. 시퀀스란 연관된 연속의 데이터를 의미하며, 시계열 데이터에 좋은 성능을 나타내는 적합한 신경망 모델이다. 음성, 문자 등 순차적인 데이터 예측에

적합한 모델이다. RNN은 시퀀스 길이에 관계없이 입력과 출력을 받아들일 수 있는 네트워크 구조이기 때문에 필요에 따라 다양하고 유연하게 구조를 만들 수 있다는 점이 RNN의 장점이다. 심층 신경망 (Deep Neural Network, DNN)의 경우 파라미터들이 모두 독립적이나, RNN의 파라미터들은 모두 공유한다. 따라서 본 논문에서는 가장 단순화된 구조로 구성되어 있는 RNN 기술과 LSTM, GRU (Gate Recurrent Unit) 기술들을 제안한 CVFE 모델에 적용하여 성능을 비교 분석하였다.

2.2.2 Long Short-Term Memory (LSTM)

LSTM은 기존 RNN 모델이 학습이 오랫동안 지속되면 초기 학습이 진행된 오래된 위치의 정보를 기억할 수 없다는 단점을 보완하여 장기, 단기 기억을 가능하게 설계한 신경망의 구조 모델이다[16]. 주로 시계열 처리나, 자연어 처리에 사용되는 모델이다. 기존의 RNN 모델에서 기억을 저장하는 메모리 셀을 추가하여 이전상태의 정보를 저장하게 한다. LSTM은 망각, 입력, 출력 게이트로 구성되어 있다. 망각 게이트는 이전 상태 정보의 저장 여부를 결정하고, 입력 게이트는 새로운 정보의 저장 여부를 결정하며, 출력 게이트는 업데이트된 셀의 출력값을 제어한다. 그림 2는 LSTM의 구조를 나타낸다.

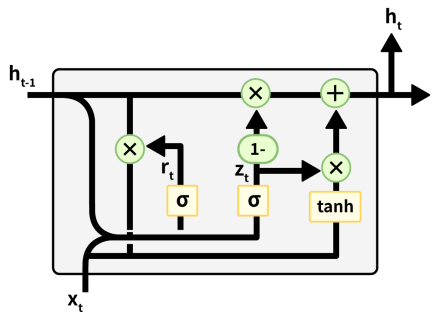


(그림 2) LSTM의 구조
(Figure 2) Structure of LSTM

그림 2의 x_t와 h_t는 각각 시간 t에서의 입력 및 히든 상태를 나타낸다. 또한 i, f 및 o는 각각 입력 게이트, 망각 게이트 및 출력 게이트를 나타낸다. LSTM의 첫 번째 단계에서 sigmoid 함수를 사용하여 삭제해야 할 정보의 양을 결정한다. 그리고 셀 상태에 새로운 정보가 저장되어야 하는지를 결정하기 위해 다른 sigmoid 함수와 *tanh* 함수를 사용한다. 이 단계를 입력 단계로 정의할 수 있다. 셀의 상태는 마지막 단계에서 업데이트되며, 출력값은 셀 상태로부터의 출력이 전달되는 최종 sigmoid 및 *tanh* 함수를 사용하여 결정된다.

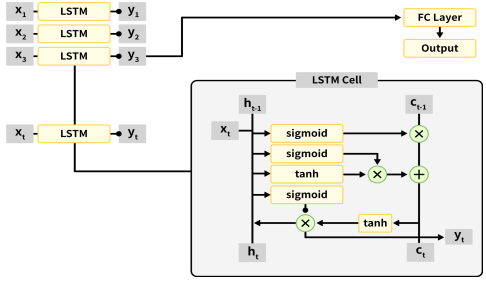
2.2.3 Gated Recurrent Unit (GRU)

GRU는 성능은 LSTM과 유사하지만 은닉 상태를 업데이트하는 연산을 단순화한 모델이다. GRU는 기존 LSTM에 비해 더 간단한 구조로 이루어져 있으며 마지막 출력값에 활성화 함수를 적용하지 않는다. 따라서 GRU는 LSTM과 유사한 성능을 갖고 있지만 학습할 파라미터가 더 적어서 학습의 속도가 빠르다. 아래 그림 3은 GRU의 구조를 나타낸다.



(그림 3) GRU의 구조
(Figure 3) Structure of GRU

그림 3과 같이 LSTM는 출력, 입력, 망각 세 가지 게이트로 구성되어 있지만 GRU에는 업데이트 게이트와 리셋 게이트의 두 가지 게이트로 구성되어 있다. 그림 3과 같이 업데이트 게이트 z_t는 다음 상태로 유지될 수 있는 정보를 결정하고 리셋 게이트 r_t는 이전 상태 정보가 GRU의 새로운 입력 정보와 결합하는 방식을 통해 결정한다. 아래 그림 4는 제안한 CVFE 모델에 LSTM 기술을 적용한 구조이다.



(그림 4) LSTM을 이용한 제안한 CVFE 모델의 구성도
(Figure 4) Diagram of proposed CVFE model using LSTM

그림 4의 y_1부터 y_t는 청크된 음성 파일의 특성인 x_1부터 x_t가 입력됐을 때 FC (Fully Connected) Layer를 거쳐서 계산된 결과 값을 나타낸다.

3. 실험 결과 및 분석

3.1 데이터 셋

본 연구에서는 음성 감정 인식 성능 시뮬레이션을 위해 IEMOCAP 음성 데이터셋을 사용하였다 [17].

(표 2) IEMOCAP 데이터셋
(Table 2) IEMOCAP Dataset

라벨	의미	데이터 갯수
ang	분노	1,103
hap	기쁨	595
exc	흥분	1,041
sad	슬픔	1,084
fru	실망	1,849
fea	무서움	40
sur	놀람	107
neu	보통	1,708
dis	역겨움	2
xxx	모름	2,507
oth	기타	3

IEMOCAP 데이터 셋은 11개의 감정으로 분류되어있으며 본 연구에선 보통 (natural), 분노 (angry), 기쁨 (happy), 슬픔 (sad) 등 네 가지 감정을 선정하여 학습 데이터로 활용하였다.

3.2 감정 인식 정확도 성능 평가

제안한 CVFE 모델의 감정 인식 정확도를 평가하기 위해, IEMOCAP의 데이터 셋을 8:1:1의 비율로 각각 훈련 데이터, 검증 데이터, 테스트 데이터로 구분하였다.

CVFE 모델에 RNN, GRU, LSTM 기술들을 적용하여 감정 인식 성능을 비교 분석하였으며 정밀도(Precision)와 재현율(Recall)을 혼동 행렬 (confusion matrix)를 사용하여 제안한 모델의 감정 인식 정확도 성능을 평가하였다.

3.2.1 정밀도 및 재현율 성능 비교 분석

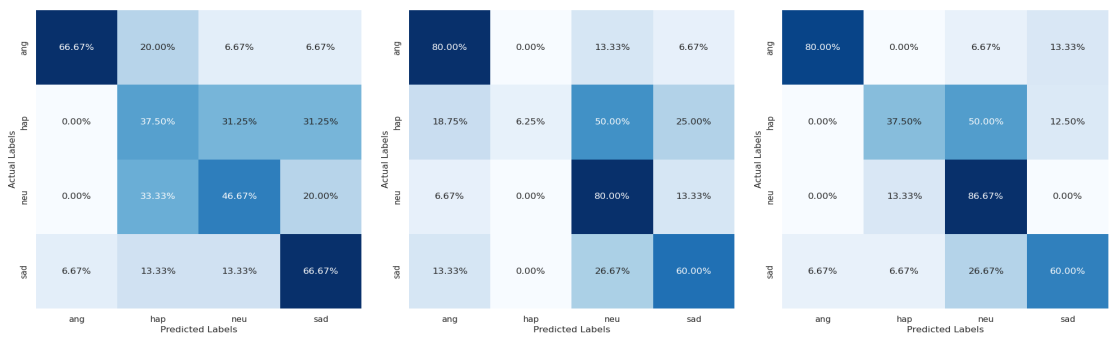
모델의 분류 성능 평가를 검증하기 위한 실험으로 사용하는 정밀도는 제안한 모델이 True라고 분류한 것 중에서 실제로 True인 값의 비율을 나타내며 아래 식 6과 같이 표현할 수 있다.

$$Precision = TP/(TP+FP) \quad (6)$$

여기서 TP(True Positive)는 실제 True인 결과를 True로 예측한 결과를 나타내며, FP(False Positive)는 실제 False인 결과를 True로 예측한 결과를 나타낸다. 반면, 재현율은 실제 True인 것 중에서 제안한 모델이 True라고 예측한 것의 비율이며 FN(False Negative)은 실제로 False인 답을 False라고 예측한 값이다.

$$Recall = TP/(TP+FN) \quad (7)$$

아래 그림 5는 제안한 CVFE 모델을 이용하여 시뮬레이션의 결과를 혼동 행렬로 나타낸 그림이다.



(a) (b) (c)

(그림 5) 정밀도와 재현율 성능의 혼동 행렬: (a) RNN, (b) LSTM, (c) GRU
(Figure 5) Confusion matrices of recall and precision performance: (a) RNN, (b) LSTM, (c) GRU

그림 5에 따르면 RNN 모델은 분노와 슬픈 감정은 높은 결과를 나타냈지만 60%대의 결과를 나타냈다. 반면 '행복'과 '보통' 감정의 정밀도와 재현율의 값이 매우 낮은 결과를 나타냈다. 하지만, 제안한 CVFE 모델에 LSTM과 GRU 기술을 적용했을 때 80%대의 높은 정밀도와 재현율을 나타냈다. 또한, '분노'와 '보통'의 감정 인식 정확도가 높았으며, 상대적으로 '슬픔'과 '행복'에서 낮은 감정 인식 정확도를 나타냈다.

아래 표3은 제안한 CVFE 모델에 LSTM 과 GRU 기술을 적용했을 때 정밀도와 재현율 성능을 비교한 표를 나타낸다.

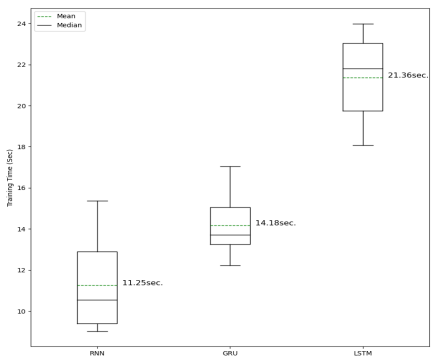
(표 3) 정밀도, 재현율, F1-Score의 성능 비교
(Table 3) Performance comparison of precision, recall and F1-Score

Model	Precision	Recall	F1-Score
RNN	0.58	0.56	0.57
GRU	0.66	0.60	0.63
LSTM	0.70	0.66	0.65

표 3에 따르면 LSTM의 정밀도, 재현율, F1-score은 각각 0.70, 0.66, 0.65로 RNN과 GRU보다 신뢰성이 가장 높다.

3.2.2 시뮬레이션 학습 시간 비교 분석

본 절에선 제안한 CVFE 모델에 RNN, GRU, LSTM 기술들을 적용하여 10번 반복하여 시뮬레이션을 진행할 때 소요되는 평균 학습시간을 비교 분석하였다.

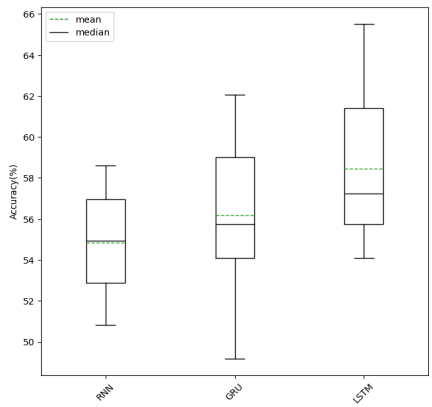


(그림 6) RNN, LSTM, GRU의 학습시간 비교
(Figure 6) Comparison of training time of RNN, LSTM, and GRU.

그림 6에 따르면, 음성 감정 인식을 위해 소요되는 학습시간은 LSTM이 평균 약 21초로 가장 높은 훈련 시간을 나타냈으며, RNN은 평균 약 11초로 가장 낮은 훈련 시간을 나타낸다. 그 이유는 하나의 LSTM 셀이 RNN 셀보다 복잡한 구조로 인해 더 많은 계산 과정이 필요하기 때문이다.

3.2.3 감정 인식 정확도 비교 분석

본 절에서는 제안한 CVFE 모델에 RNN, LSTM, GRU 기술을 적용하여 10회 반복한 시뮬레이션 결과의 평균 정확도를 비교 분석하였다.



(그림 7) RNN, LSTM, GRU의 평균 정확도 비교
(Figure 7) Comparison of average accuracy of RNN, LSTM, and GRU

그림 7은 제안한 CVFE 모델의 감정 인식 정확도를 나타낸 Box Plot 그래프이며 아래 표 4는 각 기술들의 평균 정확도를 비교한 표를 나타낸다.

(표 4) 감정 인식 정확도 성능 비교
(Table 4) Performance comparison of the accuracy of emotion recognition

Model	Accuracy	Loss
RNN	54.84%	1.011
GRU	56.19%	0.912
LSTM	58.45%	0.935

그림 7과 표 4에 따르면 제안한 CVFE 모델에 RNN 기술을 적용했을 때 54.84%로 가장 낮은 정확도를 나타냈

으며, LSTM 기술을 적용했을 때 58.45%의 가장 높은 정확도를 나타냈다. 따라서 제안한 CVFE 모델에 LSTM 기술을 적용했을 때 학습시간이 가장 많이 소요되지만 가장 높은 감정 인식 정확도를 나타낸 것을 확인하였다.

4. 결론 및 향후 과제

본 연구에선 음성 데이터로부터 정확한 감정 인식을 위해 청크 기반 감정 인식 추출 모델 (CVFE)을 제안하고 RNN, LSTM, GRU 기술들을 적용하여 시뮬레이션 학습시간과 음성 감정 인식 정확도를 비교 분석하였다. 시뮬레이션 결과, 제안한 CVFE 모델에 LSTM 기술을 적용할 때 시뮬레이션 학습시간은 가장 많이 소요되지만 가장 높은 정확도를 나타낸 것을 확인하였다.

또한 추후 CVFE 모델은 음성 연속성이 보장되어야 하는 대화형 음성에서 효과적인 성능을 낼 수 있는 모델로 향상할 수 있다. 하지만 해당 모델은 묵음이나 특정 비정상인 구간에서의 무효의 값들이 정확도를 낮추게 할 우려가 있다. 따라서 향후 연구 방향으로는 특수 음을 배제하는 데이터를 실험하고 성능 향상을 위한 연구와 시계열 데이터 배합의 규칙을 재정립하여 성능을 개선할 예정이다.

참고문헌(Reference)

[1] X. Xu, K. Meng, X. Xing and C. Chen, “Adaptive Low-resolution Palmprint Image Recognition based on Channel Attention Mechanism and Modified Deep Residual Network,” KSII Transactions on Internet and Information Systems, vol. 16, no. 3, pp. 757 - 770, 2022. https://doi.org/10.3837/tiis.2022.03.001

[2] Z. Huang, J. Li and Z. Hua, “Attention-based for Multiscale Fusion Underwater Image Enhancement,” KSII Transactions on Internet and Information Systems, vol. 16, no. 2, pp. 544-564, 2022. https://doi.org/10.3837/tiis.2022.02.010

[3] Dong C., Loy C. C., He K., Tang X, “Image Super-Resolution Using Deep Convolutional Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no.2, pp. 295-307, 2015. https://doi.org/10.1109/TPAMI.2015.2439281

[4] V. Chernykh and P. Prihodko, “Emotion Recognition from Speech with Recurrent Neural Networks,” ArXiv abs/1701.08071, 2017.

[5] R. Mu and X. Zeng, “A Review of Deep Learning Research,” KSII Transactions on Internet and Information Systems, vol. 13, no. 4, pp. 1738-1764, 2019. https://doi.org/10.3837/tiis.2019.04.001

[6] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou et al., “Speech Emotion Classification using Attention-based LSTM,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675-1685, 2019. https://doi.org/10.1109/TASLP.2019.2925934

[7] S. Mirsamadi, E. Barsoum and C. Zhang, “Automatic Speech Emotion Recognition using Recurrent Networks with Local Attention,” In Proc. 2017 IEEE ICASSP, pp. 2227-2231, 2017. https://doi.org/10.1109/ICASSP.2017.7952552

[8] X. Zhou, J. Guo and R. Bie, “Deep Learning based Affective Model for Speech Emotion Recognition,” In Proc. IEEE Conferences on UIC/ATC/ScalCom/ CBDCom/IoP/SmartWorld, Toulouse, France, pp. 841-846, 2016. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0133

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, “A Database of German Emotional Speech,” 9th European Conference on Speech Communication and Technology, Vol. 5, pp. 1517-1520, 2005. https://doi.org/10.21437/Interspeech.2005-446

[10] So-eun Park, Dae-hee Kim, “RNN-based Speech Emotion Recognition Machine Learning Algorithm,” INFORMATION AND CONTROL SYMPOSIUM, pp.152-153, 2017.

[11] Ki-duk Kim, Mi-sook Kim, “Speech Emotion Recognition through Time Series Data Classification,” Proceedings of Korea Society of Computer Information, pp.11-13, 2021. https://koreascience.kr/article/CFKO202125036398352.page

[12] Dong-jin Min, Jong-ho Won, “Time Series Feature Extraction and Performance Comparison of Recurrent Neural Network Models for Speech Emotion Recognition,” Korean Institute of Next Generation

Computing Spring Conference, pp.173-176, 2022. https://www.earticle.net/Article/A412339

[13] Seok-Pil Lee, “Feature Vectors for Speech Emotion Recognition,” INFORMATION AND CONTROL SYMPOSIUM, 226-227, 2019.

[14] Wei-Cheng Lin and Carlos Busso, “An Efficient Temporal Modeling Approach for Speech Emotion Recognition by Mapping Varied Duration Sentences into Fixed Number of Chunks,” In Interspeech 2020, Shanghai, China, pp.2322-2326, 2020. https://doi.org/10.21437/Interspeech.2020-2636

[15] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” In Proceedings of the 14th Python in Science Conference, pp. 18-25. 2015. https://doi.org/10.5281/zenodo.7746972

[16] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[17] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower et al., “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, 2008. https://sail.usc.edu/iemocap/

◐ 저 자 소 개 ◑

신 현 삼(Gil-dong Hong)
1999년 동아대학교 대학원 컴퓨터 공학과 (공학석사)
2007년 2월～현재 (주)퓨렌스 대표이사
2021년～현재 한신대학교 정보통신학과 박사과정
관심분야 : 음성처리, 음성 감정인식 등
E-mail : sam@furence.com

홍 준 기(Jun-Ki Hong)
2010년 11월 Carleton University 컴퓨터 시스템 공학과 (공학사)
2017년 2월 연세대학교 전기전자공학과 (공학박사)
2016년 8월～2017년 7월 한국정보통신기술협회(TTA) 선임연구원
2017년 8월～2020년 2월 영산대학교 전기전자공학과 조교수
2020년 3월～2023년 2월 배재대학교 컴퓨터공학과 조교수
2023년 3월～현재 공주대학교 스마트정보기술공학과 조교수
관심분야 : 인공지능, 항공체, 차세대통신 등
E-mail: jkhong@kongju.ac.kr

홍 성 찬(Sung-Chan Hong)
1983년 2월 고려대학교 통계학과(공학사)
1990년 3월 KEIO Univ. 대학원 관리공학(공학석사)
1994년 3월 KEIO Univ. 대학원 관리공학(공학박사)
2011년 1월～2012년 12월 한국인터넷정보학회 회장
2021년 2월 한신대학교 정보통신학과 교수
2021년 3월～현재 한신대학교 소프트웨어융합학부 명예교수
관심분야 : 빅데이터, 인공지능, 통계학 등
E-mail : schong@hs.ac.kr