MATE : 감정 분석을 위한 오디오-텍스트 혼합 모델

(MATE: Multimodal model using Audio and Text for Emotion recognition)

도래미솔 팀
홍성래
김태미
이 솔
김종우

카이스트 산업및시스템공학 (데이터사이언스 대학원)

INDEX

연구 배경

- 감정 정보는 의사소통의 기본 요소이며, 사람 간 상호작용에 중요한 정보임
- 지난 수년간 사용자 경험연구(User Experience: UX) 분야에서 감정 인식 자동화를 위한 다양한 방법론이 제안
- 하지만, 개인은 각기 다른 방식으로 감정을 표현하며, 감정은 시간에 따른 변화가 존재하기에 감정 인식 자동화는 여전히 도전적인 과제로 여겨짐 [1]

- 이를 극복하기 위해, 최근 **인공지능의 발전**과 더불어 **인간의 감정을 자동으로 인지하는 모델**에 대한 연구가 활발하게 이루어지고 있음

  오디오 신호에서 추출된 특징 정보를 Deep Neural Network(DNN)에 학습시켜 음성으로부터 감정을 인식하는 연구[2] 텍스트 기반의 정보로부터 감정 정보를 인식하는 방법[3] 등이 대표적

하지만, 단일 신호 기반 감정 인식 방법은 **감정 표현의 복잡성과 시간적 특징**을 고려하지 못한다는 **한계점**을 가지고 있음 [4].

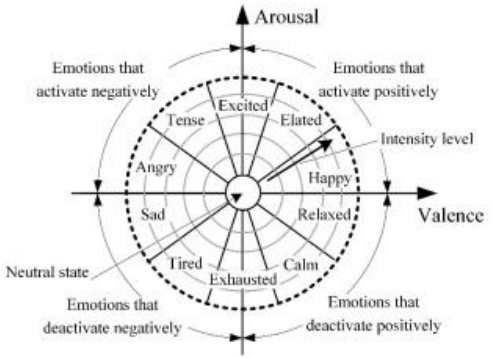


Figure 1. Russel's circumplex model of emotions.

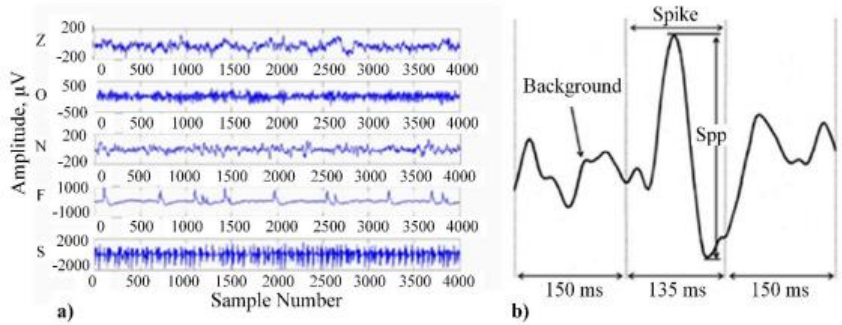


Figure 3. EEG signal: (a) example of raw data [43]; (b) peak to peak signal amplitude evaluation technique [44].

01 연구배경

문제 정의

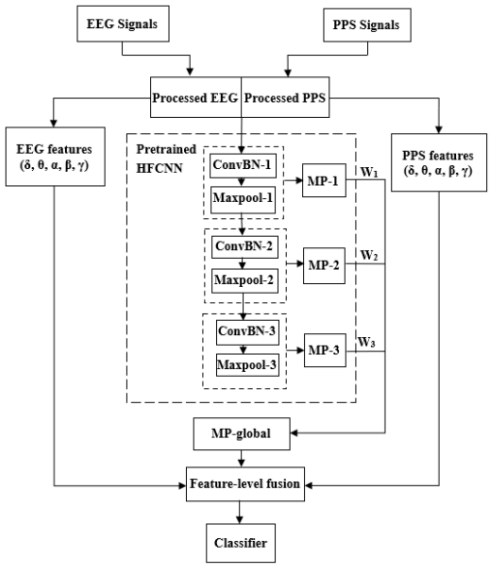
단일 신호 기반 감정 인식의 한계

- 기존 단일 신호 방법론들을 단일 데이터 자체의 정보에 의존적
- 명백한 묘사가 부족한 경우, 주어진 정보에서 정확한 감정을 분석하기 어려움
- 2개 이상의 신호를 동시에 이용하는 Multimodal Emotion Recognition (ER) 연구 필요

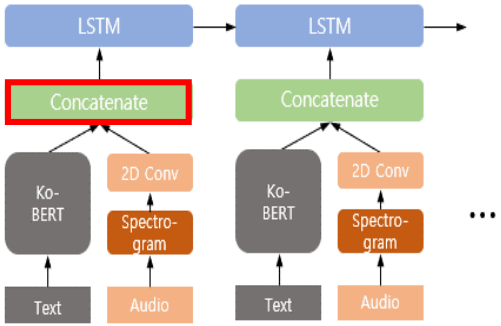
Multimodal Emotion Recognition의 Multi-path method

- [5]는 Multi-path 방식을 통해 생성된 특징 벡터를 연결하여 모델 학습에 이용
  한계점) 텍스트와 오디오의 특징을 별도로 추출, 두 데이터 간의 상호 작용 반영 어려움
- Multi-path 방식은 텍스트, 오디오 각각에 대한 상호 간 보조 정보 학습이 어려움

< Multi Path >



< Single Path >

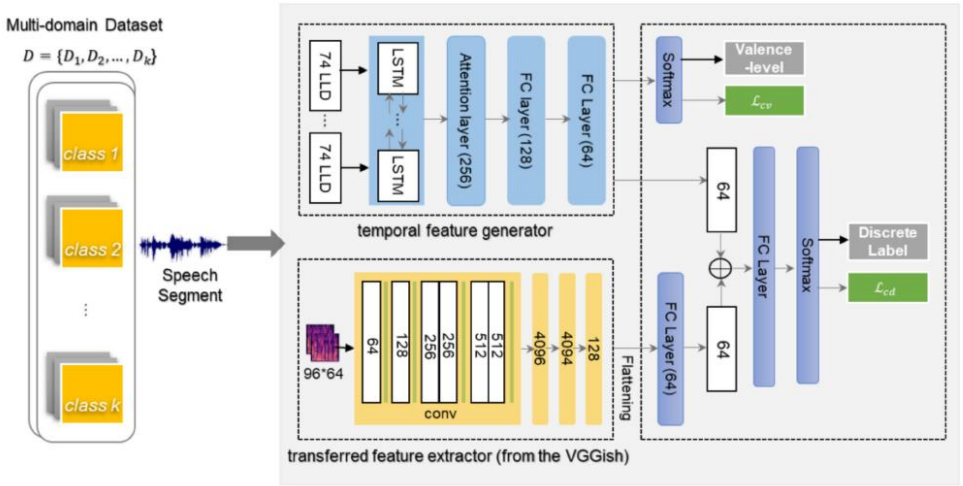


특징 학습 이전 텍스트 임베딩과 오디오 임베딩을 병합 후 LSTM Layer를 통해 하나의 특징 벡터를 생성하는 Siingle-Path 방식의 감정 분류 모델 제안

"MATE; the Multimodal model using Audio and Text for Emotion recognition"

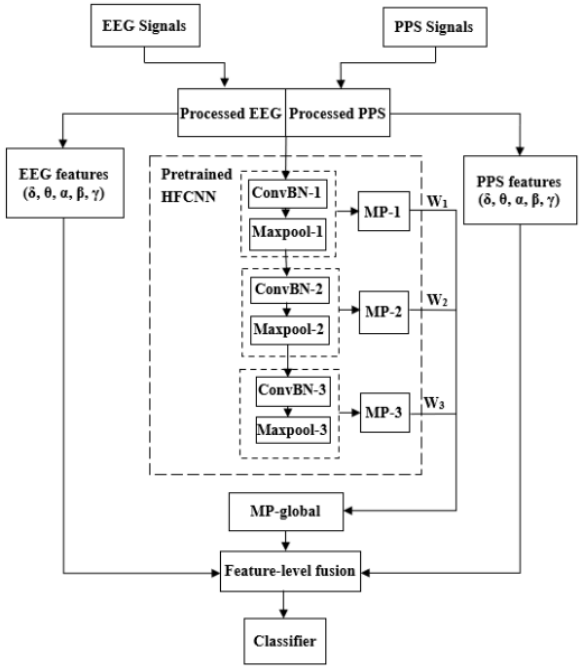
멀티모달 데이터를 이용한 감정 인식 (Multimodal Emotion Recognition)

Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets



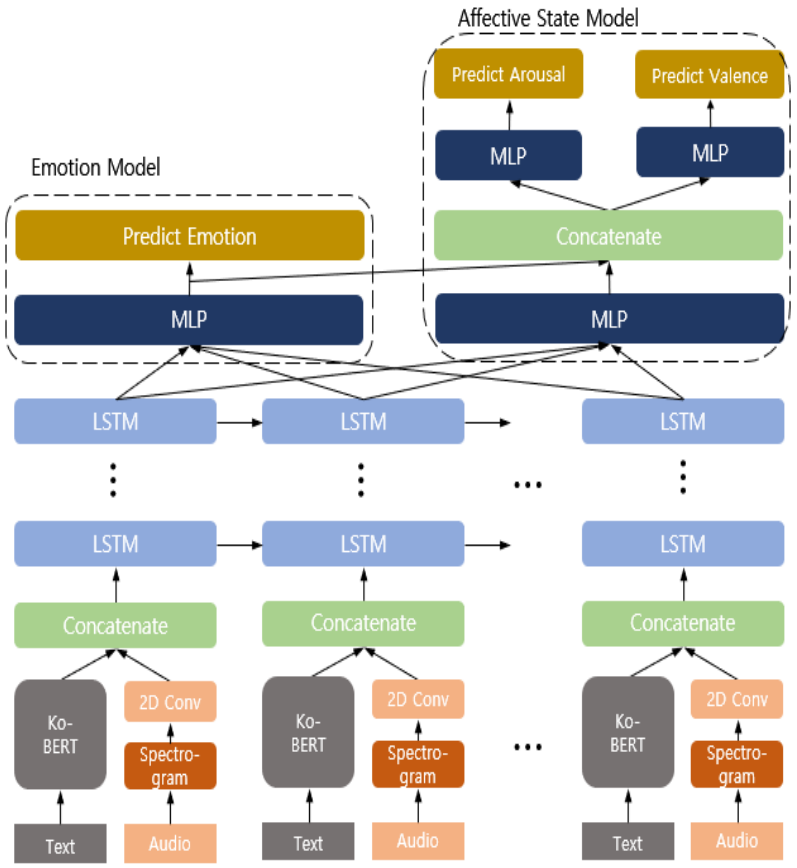
- Audio-Text Multimodal
- Multi-path
- LSTM-based text feature extractor
- VGG-like audio feature extractor

Multimodal Emotion Recognition Using a Hierarchical Fusion Convolutional Neural Network[6]



- EEG-PPS Multimodal
- Multi-path
- Hierarchical Fusion CNN model

1) Architecture



Training of Emotion and Affective State

Extracting of feature vector

Embedding of Text and Audio data

Emotion model
- 7가지 감정에 대한 확률값을 나타내 줌.
- input : (number of layer, 2048) dialog feature vector
- output : 7가지 감정의 softmax 확률을 도출.

Affective State model
- 1~5 사이의 각성도, 긍/부정도를 예측함.
- input : dialog feature vector + Emotion feature vector
- output : 4* sigmoid(x) + 1을 적용하여 1~5 사이의 각성도, 긍/부정도 정도 도출.

LSTM
- input : 1548 dimension dialog embedding
- number of layer : 8
- output : (number of layer, 2048) dialog feature vector

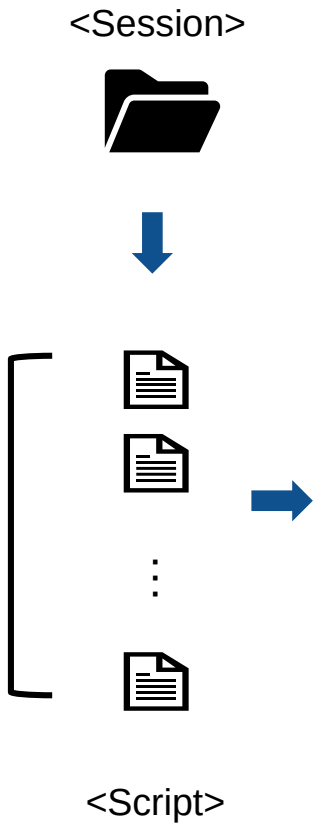
Text Data
- Kobert[7](pre-trained model)을 통해 768 dimension embedding 을 추출함.

Audio Data
- Mel-spectrogram data를 2D Conv를 통해 816 dimension embedding을 추출함.

Concatenate
- Text(768) embedding과 Audio(816) embedding을 Concatenate를 진행하여 1548 dimension dialog embedding 생성.

2) Data preprocessing

Input data



< dialog data, label >

- **청취자/발화자로 구분되는 남/여의 대화의 문자,음성** 정보가 약 40건 포함
- Emotion label
  - happy, sad, surprised, angry, neutral, disgust, fear **7가지 감정** 포함
- Affective State label
  - **감정의 각성도, 긍/부정도를 1~5의 척도**로 표현

< soft labeling >

- **10명의 평가자**의 의견에 따른 label 결정
  - 사람의 감정이 하나로 정의하기 어려운 복잡함을 가진다는 가정
  - 10명의 평가자 중 8명이 'happy'로 판단하고, 2명이 'neutral'이라고 판단했다면, 해당 데이터의 레이블 '**[0.8, 0.2, 0, 0, 0, 0, 0]**'으로 정의

3) Loss function

- **Emotion Model** : Real Emotion label distribution과 Prediction Emotion label distribution을 일치 시키기 위해 **KL-Divergence** Loss 사용.
- **Affective State Model** : 1~5 사이의 척도 차이를 알기 위해 **MSE** Loss 사용함.

1. 감정(e), 각성도(a), 긍/부정도(v)의 예측값 Loss

$$L(\hat{y}, y) = \lambda * KLDiv(\hat{e}, e) + (1 - \lambda) * (\mathrm{MSE}(\hat{a}, a) + \mathrm{MSE}(\hat{v}, v))$$

2. 'Neutral' Class가 대다수인 Class Imbalance를 위한 Class Balance Loss[8]

$$L(\hat{y}, y) = \lambda * BCE(\hat{e}, e) + (1 - \lambda) * (\mathrm{MSE}(\hat{a}, a) + \mathrm{MSE}(\hat{v}, v))$$

KL-Divergence Loss 대신 Binary Cross Entropy Loss를 사용

$$CB(\hat{y}, y) = \frac{1}{E_{n_j}} L(\hat{y}, y) = \frac{1 - \beta}{1 - \beta^{n_j}} L(\hat{y}, y)$$

n_j : Class j의 Sample 수

β : 0.99 (hyper parameter)

실험 세팅

STEP 1

파라미터 평가

- Session13~16 validation set 설정 후 파라미터 성능 비교
- 최고 성능의 파라미터 선정
- 파라미터 λ

STEP 2

최종 성능 평가

- 선정된 파라미터 모델의 5-Folds 검증 통한 최종 성능 지표 도출
- Recall, Precision, F1 score
- Concordance Correlation Coefficient
  - CCC(A): 각성도의 CCC
  - CCC(V): 긍/부정도의 CCC

모델 성능 - Speaker

[표 1 Speaker에 대한 감정, 각성도, 긍/부정도 평가 결과]

Input	λ	Recall	Precision	F1	CCC(A)	CCC(V)
Audio+Text	0.5	0.738	0.709	0.722	0.780	0.824
Audio+Text	0.66	0.748	0.719	0.733	0.745	0.860
Audio+Text	0.75	**0.759**	**0.731**	**0.744**	**0.791**	**0.869**
Audio+Text	0.8	0.696	0.670	0.682	0.783	0.803
Audio	0.66	0.489	0.466	0.477	0.540	0.588
Text	0.66	0.738	0.710	0.723	0.702	0.868

오디오 단일 신호 대비 텍스트 단일 신호 사용 시 더 높은 성능 지표 달성

오디오 또는 텍스트 단일 신호 사용 대비 오디오-텍스트 결합 시 더 높은 성능 지표 달성

모델 성능 - Listener

[표 2 Listener에 대한 감정, 각성도, 긍/부정도 평가 결과]

Input	λ	Recall	Precision	F1	CCC(A)	CCC(V)
Audio+Text	0.5	0.712	0.683	0.696	0.746	**0.880**
Audio+Text	0.66	**0.744**	**0.713**	**0.728**	**0.756**	0.865
Audio+Text	0.75	0.740	0.710	0.724	0.712	0.870
Audio+Text	0.8	0.709	0.680	0.694	0.688	0.862
Audio	0.66	0.528	0.504	0.515	0.479	0.547
Text	0.66	0.728	0.697	0.711	0.716	0.861

오디오 단일 신호 대비 텍스트 단일 신호 사용 시 더 높은 성능 지표 달성

오디오 또는 텍스트 단일 신호 사용 대비 오디오-텍스트 결합 시 더 높은 성능 지표 달성

감정이 각성도 및 긍/부정도 예측에 미치는 영향

[표 3 감정이 병합되지 않았을 때의 Speaker, Listener 성능]

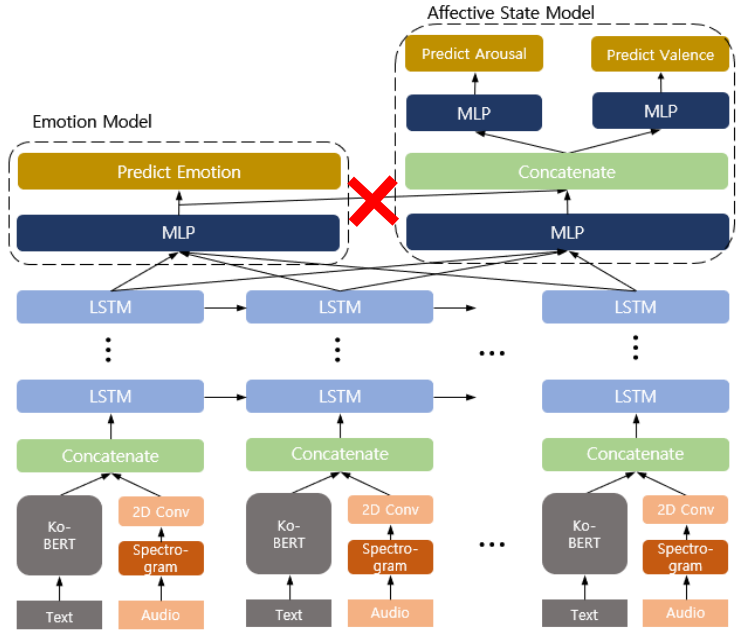
	Recall	Precision	F1	CCC(A)	CCC(V)
Speaker	0.719	0.690	0.704	0.751	0.827
Listener	0.722	0.692	0.706	0.689	0.876

※ $\lambda = 0.66$

[표 4 감정이 병합되었을 때의 Speaker, Listener 성능]

	Recall	Precision	F1	CCC(A)	CCC(V)
Speaker	0.748	0.719	0.733	0.745	0.860
Listener	0.744	0.713	0.728	0.756	0.865

※ $\lambda = 0.66$



Affective State Model에 감정 병합 시
더 높은 F1 스코어 및 높은 CCC(A)

최종 모델 5-Folds 검증 - Speaker

[표 4 Speaker 모델의 5-Folds 검증 성능]

Input	Recall	Precision	F1	CCC(A)	CCC(V)
Speaker_1	0.703	0.679	0.690	0.766	0.841
Speaker_2	0.731	0.707	0.718	0.766	0.824
Speaker_3	0.717	0.689	0.702	0.797	0.830
Speaker_4	0.759	0.731	0.744	0.791	0.869
Speaker_5	0.730	0.703	0.716	0.770	0.875
평균	**0.728**	**0.702**	**0.714**	**0.778**	**0.848**
표준편차	0.021	0.020	0.020	0.015	0.023

※ $\lambda = 0.75$

Speaker 모델 5-Folds 검증 결과 0.714 ± 0.020 의 F1 스코어 달성

0.778 ± 0.015 의 CCC(A), 0.848 ± 0.023 의 CCC(V) 달성

최종 모델 5-Folds 검증 - Listener

[표 5 Listener 모델의 5-Folds 검증 성능]

Input	Recall	Precision	F1	CCC(A)	CCC(V)
Listener_1	0.642	0.617	0.629	0.754	0.842
Listener_2	0.720	0.698	0.709	0.717	0.814
Listener_3	0.650	0.625	0.636	0.626	0.778
Listener_4	0.745	0.715	0.729	0.723	0.877
Listener_5	0.713	0.685	0.698	0.735	0.852
평균	0.694	0.668	0.680	0.711	0.833
표준편차	0.045	0.044	0.045	0.050	0.038

※ $\lambda = 0.9$, CB Loss($\beta = 0.9$)

Listener 모델 5-Folds 검증 결과 0.680 ± 0.045 의 F1 스코어 달성

0.711 ± 0.050 의 CCC(A), 0.833 ± 0.038 의 CCC(V) 달성

05 결론

기대 효과 및 추후 연구

의의

- 단일 신호가 아닌 **멀티모달** 데이터를 이용한 감정 인식 및 성능 개선
- **Single path 방식**을 통해 오디오-텍스트 간의 상호 작용을 효과적으로 반영
- **Speaker와 Listener의 감정 인식 분리**

한계점

- 감정 클래스 불균형 문제
- 오디오-텍스트의 정확한 의미 단위를 맞추지 못한 결합 방식

추후 연구

- Imbalance Sequential Data를 효과적으로 학습시키기 위한 방법론 연구
- 오디오-텍스트 결합 방식 고도화

참고 문헌

[1] C. N. Anagnostopoulos, T. Iliou and I. Giannoukos, “Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011”, Artificial Intelligence Review, Vol. 43(2), pp. 155-177, 2012.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review”, IEEE Access, Vol. 7, pp.117327-117345, 2019.

[3] A. J. AbdaouiAzé, S. Bringay and P. Poncelet, “Feel: a French expanded emotion lexicon.”, Lang Resour Eval, Vol. 51(3), pp. 833–855, 2017.

[4] J. Ma, H. Tang and W. L. Zheng, “Emotion Recognition using Multimodal Residual LSTM Network”, 27th ACM, pp. 176–183, 2019.

[5] K. J. Noh, C. Y. Jeong, J. Limm S. Chung and G. Kim, “Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets”, Sensors, Vol. 21, 2021.

[6] Zhang, Yong, Cheng Cheng, and Yidie Zhang. "Multimodal emotion recognition using a hierarchical fusion convolutional neural network." *IEEE access* 9 (2021): 7943-7951.

[7] Lee, Sangah, et al. "Kr-bert: A small-scale korean-specific language model." arXiv preprint arXiv:2008.03979 (2020).

[8] C. Yin, M. Jia and T. Y. Lin, “Class-balanced loss based on effective number of samples.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.