DBpia



# 오디오와 문자열을 활용한 다중 데이터 기반 감정 인식 모델 개발 및 검증

남영우 [O1] 이경현 [2]

[1] 성균관대학교 디지털헬스학과

[2] 인천대학교 전자공학과

spow2544@gmail.com, lkh256@gmail.com

# Development and validation of emotion recognition model using audio and text sequences: two heads are better than one

Young Woo Nam[O1] Kyung Hyun Lee[2]

[1]Department of Digital Health, SAIHST, Sungkyunkwan University

[2]Department of Electronics Engineering, Incheon National University

**요 약**

감정인식은 사람의 발화에서 감정을 예측하는 것을 목표로 한다. 초기 Multimodal Emotion recognition에서는 단순히 각 Feature를 추출한 이후 Concatenate한 이후 감정을 예측하는 방식을 사용했다. 본 논문에서는 Multimodal Emotion recognition을 위한 Multimodal Interaction Module과 Loss의 개선을 제시한다. 우리는 KEMDy19에 대한 실험을 통해 본 논문에서 제시한 방법이 감정인식 성능을 향상시킬 수 있음을 보였다. 본 논문에서 사용된 모든 코드는 1)에 공개 되어있다.

## 1. 서 론

사람은 자신의 감정과 정보를 대화를 통해 전달한다. 대화에서 정보는 명시적으로 전달하지만, 감정은 암묵적으로 전달하는 경우가 많다. 감정인식 (Emotion Recognition)은 암묵적으로 전달되는 감정을 인식하는 것은 목표로 한다. 감정인식은 Human Computer Interaction (HCI)에서 중요한 Task라고 할 수 있다[1].

사람은 대화를 할 때 몸짓, 표정, 억양 등의 여러가지 방식을 같이 사용한다. 감정인식은 Speech를 통해 감정을 인식하는 Speech Emotion Recognition, 표정을 통해 감정을 인식하는 Facial Emotion Recognition, 대화 텍스트를 통해 감정을 인식하는 Text Emotion Recognition등 여러 방식이 있다.

본 논문에서는 아래의 두가지를 제시한다.

1. RoBERTa [2]를 이용한 Text Encoder와 Wav2vec 2.0 [3]를 이용한 Audio Encoder를 결합하여 감정을 예측하는 Multimodal Speech Recognition 모델을 제시한다
2. Cross Entropy와 Cosine Similarity를 결합한 로스를 제시한다.

## 2. 관련 연구

### 2.1. Emotion Recognition

딥러닝 이전에는 speech 로부터 Filter Banks (FBanks), mel-frequency cepstral coefficients (MFCCs) 또는 hand engineered feature 를 이용하여 감정인식을 수행했다. 하지만, 최근 딥 러닝의 발전을 통해, 감정인식은 hand engineered feature 없이 원본 speech signal, 혹은 Mel-Spectrogram 를 이용하는 것으로도 좋은 성능을 내고 있다[4]. 최근, wav2vec[5]은 원본 speech signal 을 labeling 없이 이진 분류로 pretraining 한 이후, Downstream task 에 finetuning 하는 방식을 제시하였고, 그 이후 wav2vec 2.0 에서 Encoder 를 Transformer 로 대체하는 등으로 성능을 더 올릴 수 있음을 보여주었다. 또한, wav2vec 2.0 은 감정인식에서도 훌륭한 성능을 보여준다고 알려져 있다[6].

Natural Language Processing(NLP)에서는 많은 레이블이 없는 데이터셋으로 Pretrained 된 BERT [7]를 Downstream task 에 Finetuning 하는 것으로 좋은 성능을 보여주고 있다. RoBERTa 는 BERT 의 batch size 등의 hyper parameter, 학습방식을 조정함으로써 더 성능을 끌어 올릴 수 있음을 보여주었다. A. F. Adoma et.al [8]은 여러가지 pretrained model 에서 RoBERTa 가 감정인식에서 좋은 성능을 가지는 것을 보여주었다.

1) https://github.com/Mirai-Gadget-Lab/Multimodal_Emotion_Recognition

* KEMDy19 데이터셋은 https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR 에 공개 되어있다.

We thank to the National IT Industry Promotion Agency (NIPA) for the high-performance computing support program in 2021 and 2022.



www.dbpia.co.kr

### 2.2. Multimodal Emotion Recognition

감정인식은 텍스트, Speech등의 modality로 수행할 수 있다. 따라서, 2개이상의 Modality를 결합하여, 감정인식의 성능을 높이고자 하는 시도가 계속되고 있다. 초창기 Multimodal Emotion Recognition(MER)에서는 Modality 별로 feature를 추출한 이후 단순히 concatenate하는 형식으로 사용되었다 [9]. 이후, Modality간의 상호작용을 강화하기위해 Cross-Modal Attention (CMA) 모듈이 고안되었다[10-11]. CMA 모듈은 학습시에 Modality간의 참조를 통해 감정인식을 위한 Feature를 더 잘 추출하도록 설계한 모듈이다. 본 논문에서는 Srivastava.et.al [12]의 구조를 사용했다.

## 3. 방법

### 3.1. 데이터 전처리

본 논문에서는 KEMDy19 [13]를 사용했다. 텍스트 내부에 c/, n/같은 대화이외의 텍스트는 제외했다 2). 2 개이상의 감정 Label 이 붙어있는 데이터는 모델 학습에 오히려 악영향을 줄 수 있다고 판단하여 제외했다. (N: 1,298) 이후, 생체신호의 차이를 제외하면 발화, 발화 텍스트가 동일한 데이터를 제외했다. (N: 9,023 개) wav 파일이나 text 파일이 없는 경우 제외했다. (N: 7) 마지막으로, annotation file 의 wav_start, end를 차이를 기준으로 25 이상인 데이터를 제외했다. (N: 103) 이를 통해 총 10, 135 개의 데이터를 얻었다. 표 1 은 전처리 이후 데이터 분포이다.

| Emotion | Sample size |
|---|---|
| Natural | 4393 |
| Happy | 1711 |
| Angry | 1397 |
| Surpise | 858 |
| Sad | 754 |
| Fear | 690 |
| Disgust | 332 |

표 1. 전처리 이후 데이터 분포

### 3.2. 모델 구조

Srivastava.et.al은 Cross Modal Encoder를 이용하여 Modality간의 interaction 구현하고, CTC Loss를 추가함으로써, 감정인식의 성능을 개선시킬 수 있음을 보였다. 본 논문에서는 Srivastava.et.al의 모델의 구조를 사용했다. 그림 1은 본 논문에서 사용한 모델의 구조를 나타낸다.

### 3.3. Loss

Barz,B.et.al [14]은 사용가능한 데이터가 적을 때, Cross Entropy와 Cosine Loss를 결합하여 사용하면 성능의 향상을 얻을 수 있음을 보고했다. 아래는 실험에 사용한 Loss이다.

$$L = L_{cs} + \alpha \cdot L_{ce} \quad (1)$$

$L_{cs}$ 는 cosine similarity를 의미하고, $L_{ce}$ 는 cross entropy를 의미하며, $\alpha$ hyper parameter이다.

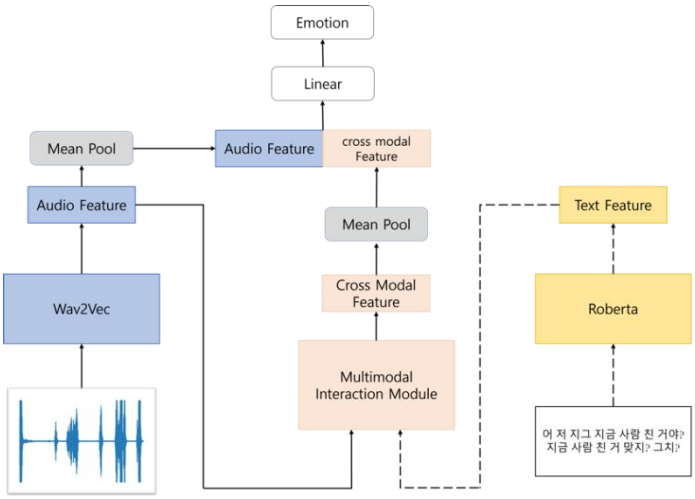


그림 1. Multimodal Interaction 모델 구조

## 4. 실험

### 4.1. 실험 환경

사전학습 모델은 Huggingface [15]로부터 다운로드 하여 사용했다. Data는 8:2 비율로 Train-Developments set과 Test set으로 분리했다. 그후 Train-Devevelopment set은 9:1로 Train, Development set으로 나누어 학습했다. Optimizer는 AdamW [16]를 사용하였고, beta_1은 0.9, beta_2는 0.999, epsilon은 1e-6. Learning rate는 5e-5를 사용했다. 텍스트만 사용한 경우 Batch size는 64, accum-grad은 1을 사용했다. 그 외의 실험에서는 Batch size 2, accum-grad는 8을 사용했다. (1)에서 사용된 $\alpha$ 는 모든 실험에 대해서 0.1을 사용했다. 본 실험은 Tesla V100 32GB 3개에서 진행되었다.

### 4.2. 실험 결과

표 2는 각 세팅 별로 Test set에 대한 Accuracy, F1-score이다. 표 2에서 CE는 cross entropy, CS는 Cosine Similarity를 의미한다. Multimodal setting에서, MMI를 사용하지 않았을 경우에는 RoBERTa, Wav2vec 2.0에서 각각 추출한 feature를 mean pooling 하고, concatenate 한 다음 Linear Layer를 통해 감정을 예측하도록 했다. CE를 로스로 사용하였을 경우, Audio, Text를 각자 사용하였을 때 보다 Multimodal로 사용하였을 때가 성능이 더 좋았다. 그러나 MMI를 사용하였을 때에는 성능이 떨어지는 것을 확인했다. CS+CE를 동시에 사용하였을 때에는, CE만을 사용했을 때와 같이, Multimodal로 사용하였을 때 성능이 더 좋았다. 그리고, MMI를 추가하였을 때, 성능이 향상되어, 모든 세팅 중에 가장 좋은 성능을 보였다.

2) c/, n/ 같은 비 발화 텍스트에 대한 annotation이 4/25일에 공개되어 추가 실험이 여의치 않아 분석에 제외함.


www.dbpia.co.kr

| Setting | Loss | Accuracy | F1-Score |
|---|---|---|---|
| Audio | CE | 0.6996 | 0.6958 |
| Text | | 0.7223 | 0.7153 |
| Audio + Text | | 0.7282 | 0.7244 |
| Audio + Text + MMI | | 0.7208 | 0.7203 |
| Audio | CS + CE | 0.7000 | 0.6903 |
| Text | | 0.7134 | 0.7094 |
| Audio + Text | | 0.7203 | 0.7149 |
| **Audio + Text + MMI** | | **0.7346** | **0.7318** |

**표 2.** Test set에 대한 Metric 결과

## 5. Conclusion

본 연구에서는 Multimodal Emotion Recognition에서 Multimodal Interaction module과 cosine similarity와 cross entropy를 결합한 로스를 사용하여 인식 성능을 높일 수 있음을 확인했다. 이후 연구에서는, 2)의 데이터를 Text Encoder에서 반영하고, KEMDy20에서 데이터를 추가하여, 감정 인식 모델의 성능을 높일 것이다. 더 나아가 생체신호 데이터를 추가 되었을 때 감정 인식 모델의 성능이 증가하는지를 연구할 것이다.

## 6. 참고문헌

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor “Emotion recognition in human-computer interaction,” IEEE Signal processing magazine, vol. 18, no. 1, pp. 32-80, 2001.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Proc. NeurIPS, 2020.

[4] A. Keesing, Y. S. Koh, and M. Witbrock, “Acoustic features and neural representations for categorical emotion recognition from speech,” in Interspeech 2021, pp. 3415-3419.

[5] S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. In Proc. of Interspeech, 2019

[6] L. Pepino, P. Riera, et al., “Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings,” in Proc. Interspeech, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL)

[8] A. F. Adoma, N. -M. Henry and W. Chen, "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition," 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2020, pp. 117-121, doi: 10.1109/ICCWAMTIP51612.2020.9317379.

[9] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” IEEE Intelligent Systems, vol. 33, no. 6, pp. 17-25, 2018.

[10] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, “Learning alignment for multimodal emotion recognition from speech,” arXiv preprint arXiv:1909.05645, 2019

[11] D. Krishna and A. Patil, “Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks.” In Interspeech 2020, pp. 4243-424

[12] Srivastava, Harshvardhan, Sreyan Ghosh, and S. Umesh. "MMER: Multimodal Multi-task learning for Emotion Recognition in Spoken Utterances." arXiv preprint arXiv:2203.16794 (2022)

[13] Noh, K.J.; Jeong, C.Y.; Lim, J.; Chung, S.; Kim, G.; Lim, J.M.; Jeong, H. Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets. Sensors 2021, 21, 1579.

[14] Barz, Bjorn, and Joachim Denzler. "Deep learning on small datasets without pre-training using cosine loss." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in EMNLP 2020, pp. 38-45

[16] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017)

2) c/, n/ 같은 비 발화 텍스트에 대한 annotation이 4/25일에 공개되어 추가 실험이 여의치 않아 분석에 제외함.


www.dbpia.co.kr