INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Audio-Visual Speech Emotion Recognition by Disentangling Emotion and Identity Attributes

Koichiro Ito, Takuya Fujioka, Qinghua Sun, and Kenji Nagamatsu

Hitachi, Ltd. R&D Group, Media Intelligent Processing Research Department, Japan
{koichiro.ito.tm, takuya.fujioka.qh, qinghua.sun.ap, kenji.nagamatsu.dm}@hitachi.com

Abstract

In this paper, we propose an audio-visual speech emotion recognition (AV-SER) that can suppress the disturbance from an identity attribute by disentangling an emotion attribute and an identity one. We developed a model that first disentangles both attributes for each modality. In order to achieve the disentanglement, we introduce a co-attention module to our model. Our model disentangles the emotion attribute by giving the identity attribute as conditional features to the module. Conversely, the identity attribute is also obtained with the emotion attribute as a condition. Our model then makes a prediction for each attribute from these disentangled features by considering both modalities. In addition, to ensure the disentanglement capacity of our model, we train the model with an identification task as the auxiliary task and an SER task as the primary task alternately, and we update only the part of parameters responsible for each task. The experimental result shows the effectiveness of our method with the wild CMU-MOSEI dataset.

Index Terms: Speech emotion recognition, Audio-visual, MTL

1. Introduction

Speech emotion recognition (SER) has been studied to make human and machine interaction more natural and to establish more engaging communication. SER has been studied using only audio [1, 2, 3, 4] or by combining other modalities [5, 6, 7, 8, 9, 10] because emotion in human speech appears as multi-modal information, such as speech content, the tone of voice, and facial expressions. In particular, the studies by [11, 12, 13] proposed SER combining tri-modal (audio, visual, and lexical) training on the CMU-MOSEI dataset [14], which is composed of speakers' videos from YouTube. Unlike a dataset recorded in a laboratory environment, speakers for such real data do not overact, and thus the clues to guess their emotion are faint. These studies demonstrated that the lexical modality mainly contributes to SER performance, but the performance estimated from the audio modality and especially from the visual one was not good. Unlike transcriptions obtained by humans, acquiring the features important for SER, such as speakers' facial expression or tone of voice, is hard from the raw sensor data in a wild dataset. However, we human can guess the person's emotion from his/her voice and appearance if we are not familiar with his/her language. In this study, we propose an SER model using only audio-visual (AV) modalities. By improving the SER performance only with the modalites, we should improve the performance when all tri-modals are combined in the future.

AV-SER is related to the AV speaker application [15, 16, 17] such as speech enhancement or separation. In these studies, the temporal synchronization between modalities is the key factor for capturing the relationship between the speaker's mouth motion and his/her voice. However, in AV-SER, emotions in each modality do not show exact temporal synchronization. A co-attention module [18], which can accept the different modality inputs to the attention module [19], has been proposed. The module has also been used in tri-modal SER studies [12, 13] as the cross-modal attention module, and the person's emotion is estimated by considering the temporal correlation of features that are extracted from different modalities. In this paper, we also follow their cross-modal attention usage.

However, estimating the emotion simply by considering the cross-modal correlation is not easy. Each modality also contains non-emotional attributes, such as the face orientation for vision and locution for audio. When we consider the cross-modal correlation, these modality-specific attributes can be ignored in the process. However, features that can identify individuals and that appear stably in both modalities every frame can be recognized for their strong correlations, and there is a concern that we can avoid extracting emotional features that are faint in the real data. To avoid this, we must disentangle both identity and emotion attributes before the cross-modal processing.

Intuitively, for example, if we consider a smiling person, estimating the person's emotion should be easy if the identity is given as the condition. Conversely, the identification should also be easy if the emotion condition is given. In this study, we introduced a cross-attributes attention (CAA) module to achieve the disentanglement, which incorporates the concept above. In this module, the emotion features are obtained by feeding the identity attribute as the conditional feature, and the identity features are also produced with the emotion attributes as a condition. In our model, both features are first disentangled in the CAA module for each modality, and then each attribute is estimated by the cross-modal processing in the cross-modal attention (CMA) module. To ensure the disentangle capacity in the CAA module, we performs training step for the identification task as the auxiliary task and the training step for SER task as the primary task alternately for each epoch, and updated only the parameters responsible for the task at each training step. In the experiment, we demonstrated that our CAA module successfully disentangled each attribute, and this also contributed to the SER performance in the wild CMU-MOSEI dataset.

2. Related Works

Emotion Recognition with auxiliary attributes: Conventional tri-modal SER studies [11, 12, 13] trained the model only with the emotion label as the supervision. However, in the acoustic field, multi-task learning (MTL) using other attributes such as emotional dimensions [2] or gender [20] has proven to be effective for SER. The MTL approach generally improves the model performance, but the features extracted through MTL tend to be entangled, so other approaches have been designed to disentangle the desired attribute's features. Facial recognition studies [21, 22] have disentangled the facial orientation and

http://dx.doi.org/10.21437/Interspeech.2021-809

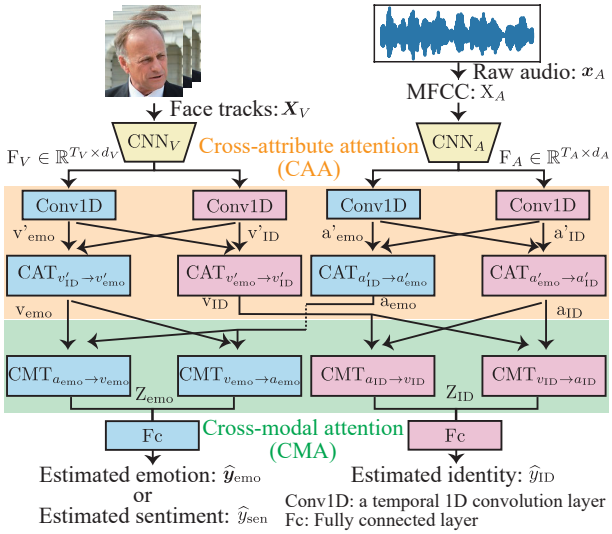
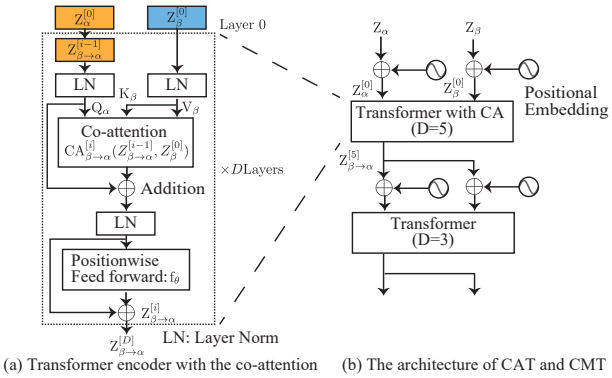


Figure 1: *Proposed Model*



(a) Transformer encoder with the co-attention (b) The architecture of CAT and CMT

Figure 2: *The modules introduced on the proposed model*

the facial action in a self-supervised manner. Studies have also achieved disentanglement by explicitly giving the auxiliary attribute as the condition. One study [23], which was inspired by [24], disentangled facial expression representations by conditioning the model giving the identity information. Another study [25] involved the same attributes, where each was disentangled by giving the features of the other attribute as the condition. Inspired by these studies, we aim to disentangle both emotion and identity attribute by giving the other attribute as the conditional feature at the CAA module for each modality.

Disentanglement in Audio-Visual speaker analysis: Disentanglement is also studied in audio-visual fields. One study [26] disentangled the identity and the speech content in a self-supervised manner. Another study [27] also disentangled the same attributes for generating talking face movies. These studies were based on confusion loss (CL) [28, 29], which is an adversarial learning method in an MTL setting where a uniform distribution is given as supervision to prevent the extracted features from containing undesired attributes. Recently, the work [30] disentangled the emotion and the identity with the same framework shown in [26] and independently extracted each attribute's features. Different from this work, our method aims to disentangle both features dependently and then estimate the speaker's emotion by considering the cross-modal interaction.

3. Method

Fig. 1 shows our model. The model requires tracked speaker's face images (face tracks) $\boldsymbol{X}_V$ and the corresponding raw audio $\boldsymbol{x}_A$ as inputs, and it estimates the speaker's identity $\hat{y}_{\mathrm{ID}}$ and the speaker's emotion $\hat{\boldsymbol{y}}_{\mathrm{emo}} \in \mathbb{R}^K$ for SER, where K is the number of the emotion class, or the speaker's sentiment $\hat{\boldsymbol{y}}_{\mathrm{sen}}$ if the task pertains to the sentiment analysis (SA). In this section, we explain our method as an SER task, though it is directly applied to the SA.

In our model, the visual representation $F_V \in \mathbb{R}^{T_V \times d_V}$ is extracted from $\boldsymbol{X}_V$ through CNN_V (CNN: a convolutional neural network). Mel-frequency cepstrum coefficients (MFCCs) are also extracted from $\boldsymbol{x}_A$ and fed to CNN_A to produce the audio representation $F_A \in \mathbb{R}^{T_A \times d_A}$. Our model is composed of two-stage structure. First, the model disentangles emotion and identity attributes with the cross-attributes attention module (CAA) in each modality. Second, the cross-modal attention module (CMA) extracts features for each attribute as the cross-modal processing from the features disentangled in the CAA. Each module in the figure is colored for describing our learning method in Sec. 3.3. In this section, we first describe the co-attention module that we utilize in both modules and then explain the details of our model.

3.1. Co-attention and the transformer encoder

Self-attention [19] has been used as co-attention [18], which extracts features considering the interaction between different data, and has also been applied to a tri-modal emotion recognition task [12, 31]. The co-attention (CA) module accepts different inputs: $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, where $T_{(\cdot)}$ and $d_{(\cdot)}$ indicate the sequential length and feature dimensions, respectively. We define queries as $Q_\alpha = X_\alpha W_{Q_\alpha}$ and keys and values as $K_\beta = X_\beta W_{K_\beta}$, $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$, $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are the parameters. The co-attention from X_β to X_α: $\mathrm{CA}_{\beta \to \alpha}(X_\alpha, X_\beta) \in \mathbb{R}^{T_\alpha \times d_v}$ is defined as follows:

$$\mathrm{CA}_{\beta \to \alpha}(X_\alpha, X_\beta) = \mathrm{softmax}(\frac{Q_\alpha K_\beta^{\mathrm{T}}}{\sqrt{d_k}})V_\beta. \tag{1}$$

Next, we describe the transformer encoder [31] in which the CA module is introduced, as shown in Fig. 2(a). The encoder has D layers, and each layer is composed of CA, layer norms (LN), and residual connections. The processing at each i layer $(i = 1, 2, \cdots, D)$ is defined as follows:

$$\begin{aligned} Z^{[0]}_{\beta \to \alpha} &= Z^{[0]}_\alpha \\ \hat{Z}^{[i]}_{\beta \to \alpha} &= \mathrm{CA}^{[i]}_{\beta \to \alpha}\left(\mathrm{LN}\left(Z^{[i-1]}_{\beta \to \alpha}\right), \mathrm{LN}\left(Z^{[0]}_\beta\right)\right) \\ &\quad + \mathrm{LN}\left(Z^{[i-1]}_{\beta \to \alpha}\right) \\ Z^{[i]}_{\beta \to \alpha} &= f_{\theta^{[i]}_{\beta \to \alpha}}\left(\hat{Z}^{[i]}_{\beta \to \alpha}\right) + \mathrm{LN}\left(\hat{Z}^{[i]}_{\beta \to \alpha}\right), \end{aligned} \tag{2}$$

where $Z^{[0]}_\alpha$ and $Z^{[0]}_\beta$ are the inputs to the module, and where f_θ has a linear parameter $W_\theta \in \mathbb{R}^{d_v \times d_\theta}$. In this processing, the output feature $Z^{[D]}_{\beta \to \alpha}$ is assumed to have meaningful information through the interaction between $Z^{[0]}_\alpha$ and $Z^{[0]}_\beta$ at the CA module for every layer. We introduce this transformer encoder to both a cross-attributes transformer encoder (CAT) and a cross-modal transformer encoder (CMT) with the positional

embedding and the ordinary transformer encoder as shown in Fig.2(b). The ordinary transformer encoder also has the same architecture but the CA module works as the self-attention. In this paper, $\mathrm{CAT}_{Z_\beta \to Z_\alpha}$ and $\mathrm{CMT}_{Z_\beta \to Z_\alpha}$ denote the operation to extract feature from Z_α attuned by Z_β using the CAT and the CMT respectively.

3.2. Proposed model

Cross-attribute attention: In the first stage, our model disentangles the emotion and identity features in each modality. Note that the disentanglement capacity in our method is guaranteed only by introducing the learning method described in Sec.3.3. The main idea to achieve the disentanglement here is to extract the target attribute's features by giving the other attribute's features as the conditions for the CAT module.

We extract each feature by running the estimation twice as follows. The visual representation F_V is fed to Conv1D modules, which independently produce the feature $v'_{\mathrm{emo}}, v'_{\mathrm{ID}} \in \mathbb{R}^{T_V \times d_V}$ for each attribute in the first estimation. In the second estimation, the emotion features in the visual modality v_{emo} are extracted from v'_{emo} and v'_{ID} through $\mathrm{CAT}_{v'_{\mathrm{ID}} \to v'_{\mathrm{emo}}}$. In this estimation, v'_{ID}, which is inputted as keys and values for the CA module in the CAT, works as the condition to produce v_{emo} from v'_{emo}. This second estimation corresponds to the operation in which we estimate the person's facial emotion when the person's identity is given. Intuitively, the emotion should be easily recognized in that condition. The visual identity feature v_{ID} is also extracted with the same operation. We explained the visual modality here as an example, but the same operation is also applied to the audio modality, and a_{emo} and a_{ID} are extracted in the same way.

Cross-modal attention: In the second stage, the emotion feature Z_{emo} and identity one Z_{ID} are respectively extracted through a cross-modal operation through the CMA as can be seen in [31]. The visual emotion features attuned with the audio ones are obtained through $\mathrm{CMT}_{a_{\mathrm{emo}} \to v_{\mathrm{emo}}}$ as $Z_{a_{\mathrm{emo}} \to v_{\mathrm{emo}}}$, and the audio emotion features attuned with the visual ones are also extracted as $Z_{v_{\mathrm{emo}} \to a_{\mathrm{emo}}}$. In this operation, the emotion features are extracted by taking the temporal correlation across the modalities, but we assume the effect of the identity attribute is reduced thanks to the disentanglement in the CAA. The emotion representation Z_{emo} is obtained as $Z_{\mathrm{emo}} = [Z_{a_{\mathrm{emo}} \to v_{\mathrm{emo}}}; Z_{v_{\mathrm{emo}} \to a_{\mathrm{emo}}}]$ with the concatenate operation, and $\hat{\boldsymbol{y}}_{\mathrm{emo}}$ is also derived. As for the identity representation Z_{ID}, the same operation is applied to v_{ID} and a_{ID}, and the identity estimation $\hat{y}_{\mathrm{ID}}$ is derived.

3.3. Learning method

This section explains the learning method that ensures the disentanglement capacity in the CAA. As shown in Fig.1, our model can be divided into three parts by color: the red, blue and yellow parts are responsible for the identification, SER, and the encoders for both tasks, respectively. We also call parameters for each part θ_{ID}, θ_{emo}, and θ_{enc}. During the training, we perform the identification as the auxiliary task and SER as the primary task alternately with the irrelevant parameters frozen for each epoch as follows.

Step 1: Identification training

In this step, we train and update only the parameters relevant to the identification θ_{ID} with the parameters for the SER θ_{emo} and θ_{enc} frozen. In this study, our primary task is the SER. Thus, we update θ_{enc} only in the Step 2 to obtain the encoder beneficial for the SER. We use cross-entropy loss (CE) as a learning objective using the identification supervision y_{ID} as follows:

$$\mathcal{L}_{\mathrm{ID}}(\hat{y}_{\mathrm{ID}}, y_{\mathrm{ID}}, \theta_{\mathrm{emo}}, \theta_{\mathrm{enc}}; \theta_{\mathrm{ID}}) = \mathrm{CE}(\hat{y}_{\mathrm{ID}}, y_{\mathrm{ID}}). \quad (3)$$

Step 2: SER training

In this step, the parameters θ_{emo} and θ_{enc} are trained to perform the SER task with θ_{ID} frozen. The dataset we used in our experiment has a multi-label for each emotion. Thus, we introduce the binary cross-entropy loss (BCE) as a learning objective using the supervision y_{emo}:

$$\mathcal{L}_{\mathrm{emo}}(\hat{y}_{\mathrm{emo}}, y_{\mathrm{emo}}, \theta_{\mathrm{ID}}; \theta_{\mathrm{enc}}, \theta_{\mathrm{emo}}) = \mathrm{BCE}(\hat{y}_{\mathrm{emo}}, y_{\mathrm{emo}}). \quad (4)$$

If the task is the SA, the loss function is defined with the supervision y_{sen} using the L1 loss instead.

We assume this learning method ensures the extraction of each attribute's feature in the CAA module because the parameters θ_{emo} and θ_{ID} are trained on each task independently.

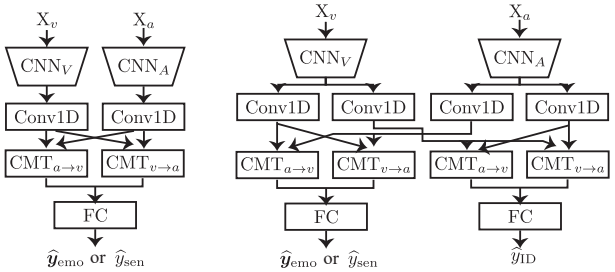
4. Experiment

4.1. Experimental Setup

Dataset: The model was trained and tested on the CMU-MOSEI dataset [14]. The CMU-MOSEI is composed of speaker video clips from YouTube with emotion and sentiment annotations. In our experiment, we utilized the raw video data instead of the pre-extracted features which is also provided. In the current released version, the dataset contains some broken samples in which the time stamp for the utterance is not correct, so the annotation is not correctly provided. We removed such samples in advance. For the supervision for both the SER and SA task, we followed the previous studies [11, 12, 13]. As for the SER, the emotions are either present or not present (binary classification), but two emotions can be present at the same time, therefore $\boldsymbol{y}_{\mathrm{emo}}$ was defined as the multi-label. As for the SA, we utilized the provided supervision $y_{\mathrm{sen}} \in [-3, 3]$ as it is. For the identity supervision y_{ID}, we utilized the YouTube video ID provided in the dataset.

Implementation: We padded or truncated each sample to keep each sample 12sec. (The average length of utterance was 7.28sec.) For the visual data, we cropped the speaker's face region using the speaker's facial landmark location provided in the dataset at 3 fps. We introduced the VGG-M architecture [32] as CNN_V and used the pre-trained model available at their website as the initial parameter. For the audio $\boldsymbol{x}_a$, we extracted 40-dimensional MFCC features with a 80-ms window size and a 20-ms hop size at a sample rate of 8 kHz. We also introduced VGG-M as CNN_A as can be seen in [26]. We extracted F_V and F_A to have $d_V = d_A = 128$ dimensional temporal features from these encoders. For the entire CMT and CAA in our model, we stacked $D = 5$ layers of transformer encoders with the co-attention and the ordinary transformer encoder ($D = 3$) as shown in Fig.2(b) and set the parameter as $d_k = d_v = d_\theta = 128$.

Evaluation Metrics: We evaluated the SER performance with weighted accuracy (WA) [33] and unweighted F1 score. We also evaluated the SA performance using the accuracy (Acc.), F1 score, and the correlation of the model's prediction with the sentiment annotation (Corr.). We introduced the equal error rate (EER) for evaluating the identification performance on the identity feature Z_{ID}. In this evaluation, we defined the score as the cosine similarity between $Z_{\mathrm{ID},i}$ and $Z_{\mathrm{ID},j}$, where i, j is the sample indices in the dataset, and the label was defined as 1 when i and j were from the same person and 0 otherwise.



(a) Baseline (b) Proposed w/o CAA

Figure 3: *Comparative Models*

4.2. Comparative Methods

The comparative methods were as follows.

Baseline: CMT was originally introduced for the tri-modal SER in the work [31]. We modified their model to accept only AV modalities to suite our experiment as shown in Fig.3(a) and treated it as the baseline. The model was trained only with the supervision of the SER or the SA.

Proposed w/o CAA: To show the effectiveness of the CAA modules in the proposed model, we evaluated the proposed model without the CAA modules as shown in Fig.3(b), toward the ablation. This model estimates each attribute independently with the two branches architecture. We trained the model in the MTL also with the identity supervision.

w/ CL: The study by [30] has proposed the two branches model for estimating each attribute trained on the framework of the CL [26, 29]. We conducted a quasi-comparison experiment by training *Proposed w/o CAA* using the CL. The difference from the original work is that they utilized 1D convolution layers for extracting temporal features, but we utilized the transformer encoder instead to suite our comparative experiment.

Proposed w/o LM: We trained our model simply as the MTL without the learning method (LM) described in Sec.3.3, which means the disentanglement capacity was not ensured in the CAA module.

Proposed w/ LM: Our method trained on the LM.

4.3. Result

The comparative results on the SER and on the SA are summarized in Table 1 and 2 respectively. The number of emotion class in the dataset was $K = 6$ but we took average from them and reported in Table 1. In our experiment, we developed the model only with the AV modalities, so the performance was worse than that in the previous tri-modal SER studies.

Proposed w/o CAA was trained on the MTL and outperformed *Baseline*. *Proposed w/ LM* was also trained on the MTL and produced better result than *Proposed w/o CAA* thanks to the larger parameters for representing the model.

In *w/ CL*, the CL was introduced to *Proposed w/o CAA* to disentangle the emotion and the identity attributes explicitly. The results show the SER performance was better than *Proposed w/o CAA*, so we determined CL was effective for the SER and SA. However, the EER suddenly degraded, which shows the identification performance was reduced.

Proposed w/ LM outperformed other methods in both the SER and SA. The architecture of the *Proposed w/ LM* was the same as that of the *Proposed w/o LM*, so the improvement could be obtained by introducing the LM described in Sec.3.3. On the other hand, EER in the Table 1 slightly degraded against *Proposed w/o LM*. We assume the degradation came from the identification training step where the encoder parameters θ_{enc} were frozen as shown in Eqn.(3), though the identification is the auxiliary task for training the CAA in our study. However, the identification performance was still better than *Proposed w/o CAA*, which indicates *Proposed w/ LM* maintained the sufficient identification capacity and this also ensured the disentanglement at the CAA module in which the identification feature was given as the conditional features. In addition, *Proposed w/ LM* was superiority to *Proposed w/o CAA* and *w/CL*, both of which extracted each attribute independently in the two branch architecture only by considering cross-modal interaction. On the other hand, our method disentangles each attributes at the CAA module in advance. We can conclude that the proposed method was effective for both the SER and the SA task thanks to the CAA module before taking cross-modal interaction at the CMA module.

Table 1: *Emotion Recognition Results: WA and F1 are for the SER. EER is for the identification (lower is better).*

Method	WA	F1	EER [%]
Baseline [31]	60.5	38.8	-
Proposed w/o CAA	60.6	38.9	14.5
w/ CL [30]	61.0	39.7	23.5
Proposed w/o LM	61.2	39.3	**14.0**
Proposed w/ LM	**61.8**	**40.4**	**14.2**

Table 2: *Sentiment Analysis Results: Acc., F1, and Corr. are for SA, EER is for the identification.*

Method	Acc.	F1	Corr.	EER
Baseline [31]	66.2	66.8	33.1	-
Proposed w/o CAA	67.4	67.2	34.3	18.3
w/ CL [30]	67.3	67.5	34.3	24.3
Proposed w/o LM	67.4	68.6	34.3	**16.8**
Proposed w/ LM	**67.6**	**68.9**	**35.9**	**16.8**

5. Conclusions

This paper proposed our AV-SER method, which disentangles emotion and identity attributes and performs SER suppressing the disturbance from the identity attribute. Our model is composed of CAA and CMA modules, which first disentangle each attribute in the CAA module in each modality and then estimate the speaker emotion by considering both modalities at the CMA module from the disentangled features. To ensure the disentanglement capacity in the CAA module, we also introduced a learning method that performs an SER task and an identification task alternately for each training epoch and in which only the parameters responsible for the task are updated for each step. In an experiment, we found that our method improved the SER and SA performance with the wild CMU-MOSEI dataset.

6. Acknowledgements

The computational resources of the AI Bridging Cloud Infrastructure provided by the National Institute of Advanced Industrial Science and Technology were used.

7. References

[1] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, 2017, pp. 2227–2231.

[2] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning." in *Proc. Interspeech*, 2017, pp. 1103–1107.

[3] S. Mao, P. Ching, and T. Lee, "Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition." in *Proc. Interspeech*, 2019, pp. 1686–1690.

[4] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. ICASSP*, 2019, pp. 2822–2826.

[5] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. Multimedia*, 2018, pp. 292–301.

[6] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. ICASSP*, 2019, pp. 6720–6724.

[7] S.-Y. Tseng, P. Georgiou, and S. Narayanan, "Multimodal embeddings from language models," *arXiv preprint arXiv:1909.04302*, 2019.

[8] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *Proc. Interspeech*, 2020, pp. 374–378.

[9] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," in *Proc. Interspeech*, 2020, pp. 3755–3759.

[10] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. ICASSP*, 2020, pp. 3507–3511.

[11] A. Shenoy and A. Sardana, "Multilogue-Net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation," in *Second Grand-Challenge and Workshop on Multimodal Language*, 2020, pp. 19–28.

[12] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Second Grand-Challenge and Workshop on Multimodal Language*, 2020, pp. 1–7.

[13] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," *arXiv preprint arXiv:2011.10652*, 2020.

[14] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. ACL*, vol. 1, 2018, pp. 2236–2246.

[15] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," *arXiv preprint arXiv:1711.08789*, 2017.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[17] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proc. ICASSP*, 2019, pp. 3965–3969.

[18] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc, CVPR*, 2019, pp. 6281–6290.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPs*, 2017, pp. 5998–6008.

[20] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning." in *Interspeech*, 2019, pp. 2803–2807.

[21] O. Wiles, A. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," *arXiv preprint arXiv:1808.06882*, 2018.

[22] Y. Li, J. Zeng, S. Shan, and X. Chen, "Self-supervised representation learning from videos for facial action unit detection," in *Proc. CVPR*, 2019, pp. 10 924–10 933.

[23] K. Ali and C. E. Hughes, "Facial expression recognition using disentangled adversarial learning," *arXiv preprint arXiv:1909.13135*, 2019.

[24] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proc. CVPR*, 2017, pp. 1415–1424.

[25] H. Wu, J. Jia, L. Xie, G. Qi, Y. Shi, and Q. Tian, "Cross-VAE: Towards disentangling expression from identity for human faces," in *Proc. ICASSP*, 2020, pp. 4087–4091.

[26] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *Proc. ICASSP*, 2020, pp. 6829–6833.

[27] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 9299–9306.

[28] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. ICCV*, 2015, pp. 4068–4076.

[29] M. Alvi, A. Zisserman, and C. Nellaaker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proc. ECCV Workshops*, 2018.

[30] R. Peri, S. Parthasarathy, C. Bradshaw, and S. Sundaram, "Disentanglement for audio-visual emotion recognition using multitask setup," *arXiv preprint arXiv:2102.06269*, 2021.

[31] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL*, 2019, p. 6558.

[32] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *In Proc. BMVC*, 2014.

[33] E. Tong, A. Zadeh, C. Jones, and L.-P. Morency, "Combating human trafficking with multimodal deep models," in *Proc. ACL*, vol. 1, 2017, pp. 1547–1556.