Exploring Multimodal Features and Fusion for Time-Continuous Prediction of Emotional Valence and Arousal

Ajit Kumar[1], Bong Jun Choi[1](✉), Sandeep Kumar Pandey[2], Sanghyeon Park[3], SeongIk Choi[3], Hanumant Singh Shekhawat[2], Wesley De Neve[3], Mukesh Saini[4], S. R. M. Prasanna[5], and Dhananjay Singh[6]

[1] Soongsil University, Seoul, South Korea
davidchoi@soongsil.ac.kr
[2] IIT Guwahati, Guwahati, Assam, India
[3] Ghent University Global Campus, Incheon, South Korea
[4] IIT Ropar, Rupnagar, Punjab, India
[5] IIT Dharwad, Dharwad, Karnataka, India
[6] ReSENSE Lab, HUFS, Seoul, South Korea

Abstract. Advances in machine learning and deep learning make it possible to detect and analyse emotion and sentiment using textual and audio-visual information at increasing levels of effectiveness. Recently, an interest has emerged to also apply these techniques for the assessment of mental health, including the detection of stress and depression. In this paper, we introduce an approach that predicts stress (emotional valence and arousal) in a time-continuous manner from audio-visual recordings, testing the effectiveness of different deep learning techniques and various features. Specifically, apart from adopting popular features (e.g., BERT, BPM, ECG, and VGGFace), we explore the use of new features, both engineered and learned, along different modalities to improve the effectiveness of time-continuous stress prediction: for video, we study the use of ResNet-50 features and the use of body and pose features through OpenPose, whereas for audio, we primarily investigate the use of Integrated Linear Prediction Residual (ILPR) features. The best result we achieved was a combined CCC value of 0.7595 and 0.3379 for the development set and the test set of MuSe-Stress 2021, respectively.

Keywords: Emotion detection · Excitation source features · Human pose · LP analysis · Multimodal fusion · Multimodal sentiment analysis

1 Introduction

Understanding emotion from different types of media content like text, audio, and video is a critical development for the early detection of mental health issues

This research was supported under the India-Korea Joint Programme of Cooperation in Science & Technology by the National Research Foundation (NRF) Korea (2020K1A3A1A68093469), the Ministry of Science and ICT (MSIT) Korea and by the Department of Biotechnology (India) (DBT/IC-12031(22)-ICD-DBT).

J.-H. Kim et al. (Eds.): IHCI 2021, LNCS 13184, pp. 729–744, 2022.
https://doi.org/10.1007/978-3-030-98404-5_65

such as depression and anxiety. The recent growth in deep learning methods has helped to improve the automated understanding of different types of media content. The research effort outlined in this paper was conducted within the context of the Multimodal Sentiment Analysis in Real-life Media (MuSe) 2021 challenge. This challenge consists of four sub-challenges, namely MuSe-Wilder, MuSe-Sent, MuSe-Stress, and MuSe-Physio [1], with these sub-challenges relying on two datasets: MuSe-CaR and U$_{LM}$-TSST [2]. The work proposed in this paper addressed the Multimodal Emotional Stress (MuSe-Stress) 2021 sub-challenge, targeting the prediction of the level of emotional arousal and valence in a time-continuous manner from audio-visual recordings. The MuSe-Stress 2021 dataset (i.e., U$_{LM}$-TSST) was distributed in two versions: (1) extracted features and (2) raw audio-visual recordings, along with a text version of the speech. We studied the results obtained by the provided baseline, leading to the observation that the extracted features had already been experimented with in different setups. Hence, we focused on engineering new features, with the goal of improving the effectiveness of emotional stress prediction.

For the video modality, we explored and experimented with various pre-trained vision models and found ResNet-50 [3] to obtain the best performance for the development test. Further, to explore novel features, we extracted new features related to human pose. In particular, we exploited the emotional attributes encoded in human body language like hand gestures, shoulder movement, neck movement, and full-body position. Extracting features from human pose was challenging due to the following reasons: (1) hand gesture identification: all participants have a sensor in one hand, which makes it difficult for a model to identify both hands (often only one hand was detected), (2) foot gesture identification: the video was recorded at a certain distance, with many frames having missing lower body parts, and (3) natural movement: all participants were equipped with sensors, with these sensors restricting the natural movement of body parts, thus limiting the availability of pose and associated emotional information. Some of these challenges were resolved through the use of OpenPose [4], YOLO [5], and pre-processing (e.g., resizing of the bounding box of a hand). We observed that human pose is able to provide good features for detecting human emotions. Still, to solve the aforementioned challenges in a more effective way, other approaches can be leveraged to obtain better features from human gestures.

For the audio modality, we experimented with the provided features and also investigated new features, namely Integrated Linear Prediction Residual (ILPR) features [6]. Our final audio predictions were based on the combination of eGeMAPS [7] and ILPR features. For the text modality, we experimented with BERT [8] features. Still, from the baseline and based on our experiments, we learned that text features generally degrade the performance when combining them with other features. We largely used the model and code given as the baseline to implement our deep learning approach, tuning various parameters like the network size and the learning rate. Since early fusion (concatenation of features before model training) was not resulting in a good performance, we adopted the late fusion approach used by the baseline.

Our best performing combination achieved CCC values of (0.7549, 0.7640), outperforming the baseline result (0.5043, 0.6966) for the development set (for arousal and valence, respectively). However, our best performing combination (0.2297, 0.4158) did not do well on the internal test set of MuSe-Stress 2021, compared to the baseline test result (0.4562, 0.5614). We plan to further mitigate this performance gap in future work.

The remainder of this paper is organized as follows. Section 2 discusses the features and the model used by our approach. Section 3 describes our experiments and the results obtained. Finally, Sect. 4 concludes the paper.

2 Features and Model

2.1 Audio

Speech, which is the preferred form of communication between individuals, carries a substantial amount of primary information (e.g., the intended message) and secondary information (e.g., emotion, speaker identity, and gender). Speech production characteristics are different for neutral speech and emotionally charged speech. This difference in speech production characteristics under different emotion scenarios is reflected in terms of changes in the vocal tract system and the excitation source [6]. This motivated us to explore information about both the excitation source and the vocal tract for feature extraction for the purpose of time-continuous prediction of emotional valence and arousal.

Excitation Source Features for Audio. The task of separation of the vocal tract and excitation source information from a given speech signal has been explored in the literature [9–11]. Researchers have suggested the use of Linear Prediction (LP) analysis, with proper order selection [12] to model the vocal tract and excitation source information. The LP Coefficients obtained from LP analysis represent the vocal tract information, and the LP residual signal obtained using the inverse filter represents the excitation source information. For a detailed overview of the Integrated Linear Prediction Residual (ILPR) signal extraction process, please refer to [13]. A brief description regarding the extraction of the ILPR signal is given below.

Given a speech signal $s[n]$, where n represents the sample index and $n = 0, 1, \ldots$, the first stage involved is using a pre-emphasis filtering to enhance the high-frequency components of the speech signal. The filtered signal $s_e[n]$ is then used for further LP analysis. The LP analysis works by predicting the present speech sample $\hat{s_e}[n]$ based on a combination of p past speech samples $(s[n-1], \ldots, s[n-p])$. This combination of p past samples can be mathematically represented as

$$\hat{s_e}[n] = -\sum_{k=1}^{p} c_k s_e[n-k], \tag{1}$$

where c_k represents the Linear Prediction Coefficients (LPCs). The LP residual (LPR) signal $e[n]$ can then be computed as follows:

$$e[n] = s_e[n] + \sum_{k=1}^{p} c_k s_e[n-k]. \tag{2}$$

The LPCs are computed using the Levinson-Durbin Algorithm [14]. An inverse filter $I_{lp}(z)$ is realized using the computed LPCs. The transfer function of the inverse filter in the z-domain is given as follows:

$$I_{lp}(z) = 1 + \sum_{k=1}^{p} c_k z^{-k}. \tag{3}$$

Finally, the ILPR signal is obtained by passing the original speech signal $s[n]$ through the inverse filter $I_{lp}(z)$. The resulting ILPR signal can then be further used to compute speech features.

eGeMAPS Features. eGeMAPS features for the audio modality are part of the baseline feature set. In accordance with the time step defined in the baseline, eGeMAPS features of 88 dimensions are computed from the ILPR signal at a time step of 500 ms. The ILPR eGeMAP and the Speech signal eGeMAPS are concatenated to form a combined feature vector of dimensionality 164. The ILPR eGeMAP represents information from the excitation source, whereas the speech eGeMAP represents vocal tract characteristics. This combined feature vector is used in our work to model the emotional information from the audio modality.

2.2 Video

ResNet. Computer vision applications such as object detection and object recognition are actively exploring the use of Convolutional Neural Networks (CNNs), given the ability of these artificial neural networks to identify features of interest without human intervention. Moreover, multi-layered CNNs like VGG can obtain a high effectiveness for a multitude of prediction tasks, including the task of face recognition [15].

In general, CNNs perform better as the number of layers increases [16]. Indeed, a higher number of layers allows a neural network to learn more complex features, eventually enabling the neural network to solve more complex tasks. However, one drawback of deep neural networks is the occurrence of vanishing gradients during backpropagation, making it difficult to train multi-layered neural networks. The issue of vanishing gradients refers to the phenomenon where the gradients of the loss function get smaller as they are propagated towards the initial network layers, given the application of the chain rule.

Residual Networks (ResNets) [3] were proposed in 2016 to mitigate the vanishing gradient problem. A ResNet is composed of a series of convolutional layers with residual blocks. These blocks have skip connections that allow gradients to flow directly from the later convolutional layers to the initial convolutional layers. For the MuSe-Stress 2021 sub-challenge, we leveraged ResNet-50 to extract features from the video modality, obtaining these features from the last convolutional layer present in this network architecture. Note that ResNet-50 is a network deeper than VGG-16 [15], where the latter is used as a baseline architecture by the organizers of the Muse-Stress 2021 sub-challenge. Furthermore, ResNet-50 outputs 2048-D features, whereas VGG-16 produces 512-D features.

The inputs used for feature extraction are the facial images cropped directly from the raw video using MTCNN [17]. Due to variations in size, all facial images were re-sized to a resolution of 224 × 224 before being fed into a ResNet-50 pre-trained on VGGFace2 (no fine-tuning was used). VGGFace2 [18], a further development of VGGFace [19], is a large-scale dataset for recognizing faces, consisting of 3.31 million images of 9,131 identities. In what follows, the features (activation vectors) extracted through ResNet-50, pre-trained on VGGFace2, will be referred to as *RN50*.

OpenPose. To extract additional features from the video modality, we made use of OpenPose, an open-source system for multi-person 2-D pose detection in real time [4]. The OpenPose approach, which relies on the use of stacked CNNs, consists of two major parts: while one part predicts 2-D confidence maps for keypoints of interest (e.g., body, foot, hand, and facial keypoints), the other part predicts so-called Part Affinity Fields (PAFs), with a PAF referring to a set of 2-D vector fields that encode the location and orientation of limbs over the image domain [4,20].

Among the different kinds of features that can be extracted by OpenPose, we decided to work with facial features and body&foot features. Throughout the remainder of this paper, these features will be further referred to as *OFF* and *OBF*, respectively. For *OFF*, 70 2-D keypoints were extracted, whereas for *OBF*, 25 2-D keypoints were extracted[1]. The 70 *OFF* and 25 *OBF* keypoints are shown in Fig. 1.

As *OFF* and *OBF* aim at representing different types of visual information, the inputs used to extract the aforementioned features are different. In particular, the inputs used for facial feature extraction are the same facial images used for ResNet-50 feature extraction. Furthermore, unlike the cropped facial images used for obtaining *OFF*, *OBF* is created using the corresponding full-resolution images. Given these different inputs, we made use of two different pre-trained models. First, for OBF, we made use of the body25 model, which has been pre-trained on a combination of the MPII dataset and a subset of foot instances taken from the COCO dataset [21], and where the latter was labeled using the Clickwork platform [21]. Second, for OFF, we made use of a face model that has been pre-trained on facial keypoints taken from the COCO dataset [4].

[1] The 25 *OBF* keypoints are Nose, Neck, R/L Shoulders, R/L Elbows, R/L Wrists, MidHip, R/L Hips, R/L Knees, R/L Ankles, R/L Eyes, R/L Ears, R/L BigToes, R/L SmallToes, R/L Heels, and Background (R/L stands for Right/Left).

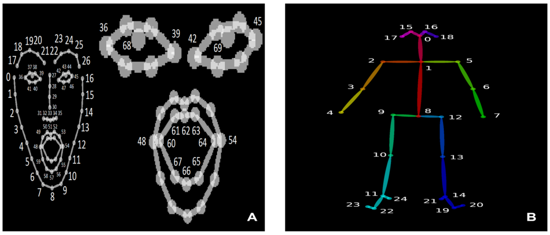

Fig. 1. Image A shows the 70 facial keypoints extracted by OpenPose. Image B shows the 25 body&foot keypoints identified by OpenPose.

Besides the use of facial features and body&foot features, we also investigated the extraction of hand features. The U_{LM}-TSST dataset is composed of videos that feature free-talking participants in a stressful situation, and with these participants making use of minimal body movements. However, while talking, the participants are still freely moving their hands around. As such, we hypothesize that these hand gestures are representative of the emotional state of the participants (i.e., we hypothesize that these hand movements convey emotional information).

OpenPose can identify 21 hand keypoints, making available a model that has been pre-trained on the COCO dataset. However, the extraction of hand features using OpenPose is still a work in progress due to the limitations discussed below.

- OpenPose is unable to identify keypoints from full-resolution images properly. This is for instance illustrated in Fig. 2, where most of the keypoints are wrongly identified. Hence, we decided to crop hands before doing keypoint recognition using You Only Look Once version 3 (YOLOv3) [5].
- Using the cropped hand images as input for feature extraction purposes also comes with limitations. Indeed, OpenPose can only identify keypoints from one hand at a time, as depicted in Fig. 3B. Furthermore, difficulties are also encountered when an object (e.g., a sensor) is attached to a hand, as illustrated in Fig. 3C.

YOLOv3, which is pre-trained on the CMU Hand DB dataset [22], was used to detect and crop hands from full-resolution images. However, it is still necessary to choose a proper bounding box size to avoid having cropped images that contain both hands. In addition, we believe the detection problems that arise due to a hand holding an object can be addressed through the use of skin masking for removing this object [23] (alternatively, a future version of the U_{LM}-TSST dataset may avoid the presence of hands with sensors attached altogether). Once the above-mentioned limitations are overcome, it should be possible to leverage the extracted hand gesture information to better predict valence and arousal values in a time-continuous manner.

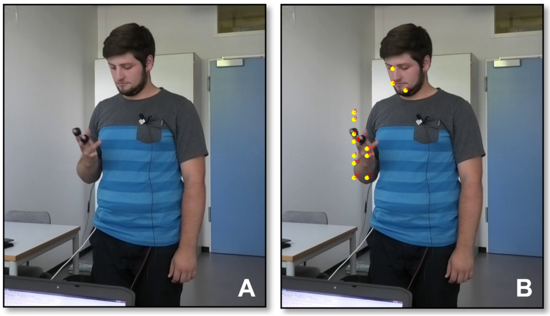

Fig. 2. Image B shows the hand keypoints recognized by OpenPose in the full-resolution Image A.

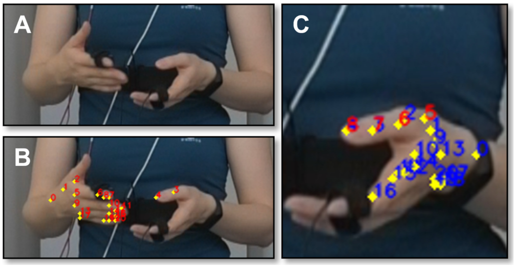

Fig. 3. Image A shows a hand image cropped from a full-resolution image. Image B shows the keypoints identified in Image A using OpenPose. Image C visualizes the keypoints found when an object is attached to a hand.

2.3 Text

The Bidirectional Encoder Representations from Transformers (BERT) [8] features were provided for the text modality. We observed from the baseline results that these features do not perform well in a standalone setting, even degrading the performance when used in conjunction with other modalities (i.e., audio and video). We explored the possible causes for the low performance of BERT features. The first cause is related to the breaking of text based on segment length rather than the end of a sentence, with the new model SentenceBERT [24] also having this problem. From the literature, one can learn that BERT features work well with full sentences, with full sentences helping to capture the context in which a word is used. The second cause is related to human psychology: in a stressful setting, humans are less likely to speak, resulting in lots of blank segments. We believe that these segments must have emotional information but we could not verify this using the other modalities as we aimed at improving the performance. We can take this as another direction for future research.

2.4 Fusion

Humans are essentially complex multimedia systems. They have multiple senses and express their behavior in terms of multiple modalities, such as voice, facial appearance, activity, and text. With the rapid development of sensing technologies, it is easy to capture these modalities automatically. There are two main techniques for fusing multimodal information: early fusion and late fusion [25]. In early fusion, we combine the features from multiple modalities at an early stage and train a classifier on top of it. Various early-fusion techniques have been explored for early fusion, particularly for image analysis. It has been observed that early fusion is effective in scenarios where the spatio-temporal organization of the data is similar (e.g., RGB images and thermal images) [26]. This is probably because, at the initial stage, the feature maps still contain local structure.

In our case, the modalities used (text, audio, and video) have different spatio-temporal structures. In addition, a single classifier would not be able to learn the class distributions of the individual modalities effectively. Therefore, we are inclined to make use of late fusion. Yet, instead of using hand-crafted weights to fuse the decisions from individual modalities, we feed the late multimodal features to an LSTM-RNN and let it learn the fusion weights automatically by looking at the data [20]. The LSTM does not only exploit the complementary information from multiple modalities, but it also implements a self-attention mechanism in the given time-series data. We find that such mid-level fusion gives the best results on the given dataset.

3 Experiments and Results

3.1 Experimental Setup

Video. All experiments with features extracted from the video modality (i.e., *RN50*, *OBF*, and *OFF*) were performed using the baseline LSTM-RNN model[2]. In particular, a bi-directional LSTM-RNN was used and trained for 100 epochs, relying on early stopping with a patience of 15 epochs. The number of seeds was set to 10, saving the predictions made by the best-performing model. The features and labels were segmented as stated in [1], using a window size and a hop size of 300 steps (150 s) and 50 steps (25 s), respectively.

A grid search was performed to investigate the optimum combination of different hyperparameter settings, taking into account the normalization of input features (n), the hidden state dimension (h), the number of LSTM-RNN layers (r), and the learning rate (lr). The values tested for each hyperparameter are as follows: $n = \{\text{True, False}\}$, $h = \{8, 32, 64, 128\}$, $r = \{1, 4, 5, 9\}$, and $lr = \{2\text{e}{-5}, 5\text{e}{-5}, 0.002, 0.2\}$.

To mitigate issues in terms of overfitting, three different methods were investigated in a later stage of our experimentation: dropout, L_2 regularisation, and data augmentation. Similar to the previously described hyperparameter tuning

[2] https://github.com/lstappen/MuSe2021.

Table 1. Valence and arousal results obtained for hyperparameter tuning when making use of the *RN50* features and lr = 0.002.

Hyperparameter tuning											
Valence						Arousal					
h	*r*	*n*	*d*	*L2*	CCC	*h*	*r*	*n*	*d*	*L2*	CCC
64	4	False	0.0	0.0	0.3310	64	4	False	0.0	0.0	0.2022
128	5	False	0.0	0.0	0.6108	64	4	True	0.0	0.1	0.0036
128	5	False	0.5	0.0	0.6056	64	4	True	0.0	1e−6	0.2954
128	5	False	0.7	0.0	0.5078	64	4	True	0.0	1e−8	0.3339
128	5	False	0.5	1e−4	0.5687	64	4	True	0.5	1e−8	0.3397
128	5	False	0.7	1e−4	0.5875	64	4	True	0.7	1e−8	0.3955
128	5	False	0.5	1e−6	0.6198	16	4	True	0.0	0.0	0.4088
128	5	False	0.7	1e−6	0.5725	16	4	True	0.7	1e−8	0.3295
128	5	False	0.5	1e−8	0.6249	64	4	True	0.2	0.0	0.3320
128	5	False	0.7	1e−8	0.5802	64	2	True	0.0	0.0	0.2241
128	5	False	0.3	1e−8	**0.6337**	64	4	True	0.0	0.0	**0.4368**

approach, the impact of different values for dropout $(d) = \{0.2, 0.3, 0.5, 0.6, 0.7\}$ and L_2 regularisation $(L2) = \{0, 1\text{e}{-}8, 1\text{e}{-}6, 1\text{e}{-}4, 0.1\}$ was investigated. Data augmentation was applied to the facial images through random horizontal flipping (flipping probability used: 0.5), leading to an increase in dataset size. For each type of visual feature used, the best predictions obtained for the valence and arousal dimensions were then passed onto the late fusion stage of the proposed approach.

Late Fusion. All experiments with late fusion made use of an LSTM-RNN model with the following default hyperparameter values: $\{h, r, lr\} = \{32, 2, 0.001\}$. The exception is the late fusion of *OFF* and other modalities (third submission), for which $h = 64$.

3.2 Results and Discussion

Video. Table 1 shows the CCC values obtained when predicting continuous arousal and valence values using the *RN50* features for the development set and for different hyperparameter settings. The best results achieved are indicated in bold. Given the different hyperparameter settings, we were able to obtain the best results for valence when making use of 128-D hidden states, a learning rate of 0.00005, a dropout value of 0.3, and an L_2 penalty of 1e–8. In addition, we obtained the best results for arousal when using 64-D hidden states, a learning rate of 0.002, and feature normalization, not using dropout and not using L_2 regularisation. In summary, the highest CCC value achieved for the development set is 0.6337 for valence and 0.4368 for arousal.

Table 2. Impact of data augmentation on valence and arousal when making use of the *RN50* features. CCC$_i$ and CCC$_f$ refer to the CCC values obtained, without and with data augmentation, respectively.

Data augmentation															
Valence								Arousal							
lr	*h*	*r*	*n*	*d*	*L2*	CCC$_i$	CCC$_f$	*lr*	*h*	*r*	*n*	*d*	*L2*	CCC$_i$	CCC$_f$
0.00200	64	4	False	0.0	0.0	**0.3310**	**0.5089**	0.002	64	4	False	0.0	0.0	0.2459	0.3203
0.00005	128	5	False	0.0	0.0	0.6108	0.5027	0.002	64	4	True	0.0	0.0	**0.4193**	**0.3413**
0.00005	128	5	False	0.3	0.0	0.6253	0.5583	0.002	64	4	True	0.5	0.0	0.3187	0.3468
0.00005	128	5	False	0.5	0.0	0.5904	0.5919	0.002	64	4	True	0.0	1e−4	0.1794	0.3094
0.00005	128	5	False	0.7	0.0	0.5717	0.5682	0.002	64	4	True	0.5	1e−4	0.2517	0.2926
–	–	–	–	–	–	–	–	0.002	64	4	True	0.7	1e−4	0.3200	0.3277
–	–	–	–	–	–	–	–	0.002	16	1	True	0.5	0.0	0.1429	0.1830
–	–	–	–	–	–	–	–	0.002	16	4	True	0.0	0.0	0.1387	0.1083

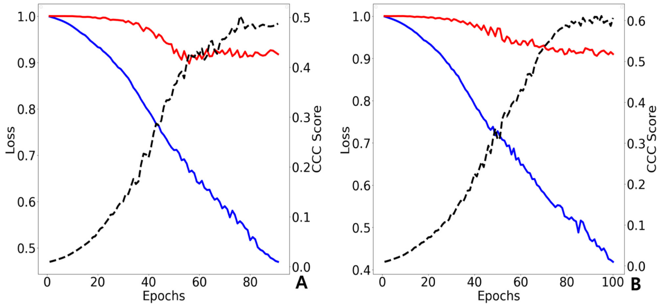

Fig. 4. Valence learning curves: (A) after data augmentation and (B) before data augmentation. The blue, red, and black dotted lines represent train loss, validation loss, and development CCC score, respectively. (Color figure online)

Additionally, Table 2 shows how data augmentation affects the predictions made. Given the valence results presented in Table 2, we can observe that data augmentation often helps in getting better predictions along this dimension when not making use of hyperparameter tuning (row 1); however, the CCC values obtained after data augmentation and hyperparameter tuning are lower than the CCC values obtained before data augmentation but after hyperparameter tuning (rows 2–5). Furthermore, the arousal results presented in Table 2 also show that data augmentation often helps in getting better predictions along this dimension; however, the highest CCC value obtained after data augmentation is still lower than the highest CCC value obtained before data augmentation.

Furthermore, Fig. 4 shows how the training and validation loss change as a function of model training, depicting the training loss using a blue line and the validation loss using a red line. We can observe that the validation loss in Fig. 4(A) follows the training loss more closely than in Fig. 4(B). However, the highest CCC value obtained is much higher in Fig. 4(B). For arousal, we can observe trends similar to the trends observed for valence, as illustrated by Fig. 5.

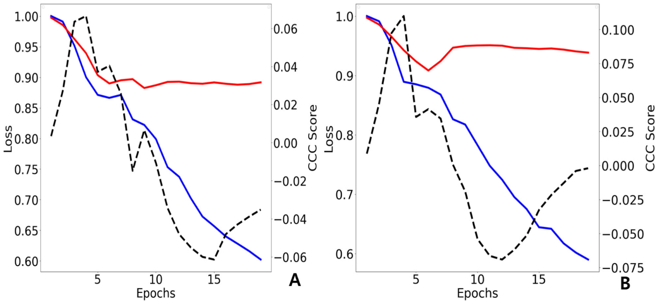

Fig. 5. Arousal learning curves: (A) after data augmentation and (B) before data augmentation. The blue, red, and black dotted lines represent train loss, validation loss, and development CCC score, respectively. (Color figure online)

Table 3. Valence and arousal results obtained for our five submissions, levering late fusion of different modalities (A for audio, V for video, and T for text). The best CCC values obtained for the development set and the test set are highlighted in bold.

Submission	Features	Valence	Arousal	Combined
		dev/test	dev/test	dev/test
1	Best A + Best V Best A + Best V + Best T	–/– 0.6916/0.354	0.6324/0.1798 –/–	0.6620/0.2669
2	*OFF*	0.6024/0.2953	0.5203/0.0403	0.5613/0.1678
3	Best A + Best V + *OFF* Best A + Best V + Best T + *OFF*	–/– 0.7366/**0.4461**	**0.7553**/0.1330 –/–	0.7456/0.2896
4	Best A + Best V + *OBF* Best A + Best V + Best T + *OBF*	–/– 0.7393/0.4113	0.7017/0.2159 –/–	0.7205/0.3146
5	Best A + Best V + *OFF* + *OBF*	**0.7640**/0.4158	0.7549/**0.2297**	**0.7595**/**0.3228**

In particular, we can again observe that the validation loss follows the training loss more closely in Fig. 5(A) than in Fig. 5(B). However, the highest CCC value in Fig. 5(A) is still lower than the highest CCC value that can be found in Fig. 5(B).

Given the experimental results obtained, we can conclude that data augmentation can help in preventing overfitting. Nevertheless, for the MuSe-Stress 2021 sub-challenge, we decided not to make use of data augmentation. Indeed, when hyperparameter tuning is in place, most CCC values obtained after data augmentation are then lower than the corresponding CCC values obtained before data augmentation, for both valence and arousal. Furthermore, since we only paid a limited amount of attention to the use of data augmentation for mitigating the risk of overfitting, we believe future work could still explore the use of other types of data augmentation, such as random vertical flipping and random cropping, as well as an optimization strategy that makes it possible to effectively combine hyperparameter tuning and data augmentation.

Submission Results. Table 3 shows the results obtained for our five submissions. All submissions leveraged late fusion, except for the second submission, which obtained the predictions when only making use of *OFF*. The best A (audio), V (video), and T (text) values represent the best development scores obtained for each modality, for which the feature and hyperparameter combinations used are summarized below.

- Best A: Combining *ILPR eGeMAPS* features with an LSTM-RNN ([h, r, lr] = [64, 4, 0.002]) yielded a development score of 0.5632 and 0.4841 for valence and arousal, respectively.
- Best V: The models used for predicting valence and arousal adopted different settings.
  - Valence: Combining *RN50* features with an LSTM-RNN ([h, r, lr] = [128, 5, 5e−5]) yielded a development score of 0.6253.
  - Arousal: Combining normalized *RN50* features with an LSTM-RNN ([h, r, lr] = [128, 9, 0.002]) yielded a development score of 0.4399.
- Best T: Combining *BERT* features with an LSTM-RNN ([h, r, lr] = [64, 4, 0.002]) yielded a development score of 0.3626 and 0.2308 for valence and arousal, respectively.

For each submission, the combination of features submitted for valence and arousal typically differed as it is not necessary to use identical modalities to predict valence and arousal. Our third submission obtained the highest test score for valence, namely 0.4461. Furthermore, our fifth submission obtained the highest test score for arousal, namely 0.2297, and the highest combined score, namely 0.3228.

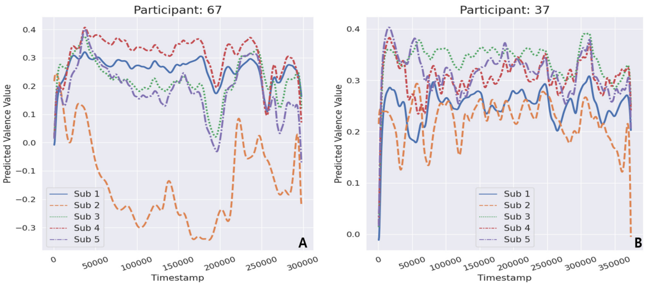

Fig. 6. Late fusion results for valence: (A) Participant 67 and (B) Participant 37.

Discussion of Submission Results. Figure 6 and Fig. 7 show the valence and arousal values obtained for a number of participants of interest (i.e., participants showing a clear contradicting result) when leveraging late fusion, combining the speech, text, and video modalities. The labels Sub 1 to Sub 5 refer to the corresponding submissions listed in Table 3.

For Fig. 6(A), we can observe that all of the curves follow a similar trend, except for Sub 2. Different from Fig. 6(A), we can observe that all of the curves in Fig. 6(B) follow a similar trend. Also, using Table 3 to examine Sub 2 in more

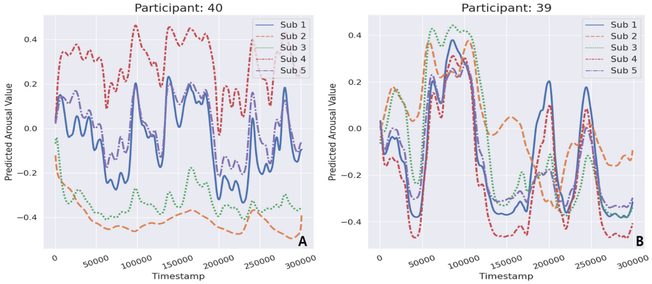

Fig. 7. Late fusion results for arousal: (A) Participant 40 and (B) Participant 39.

detail, which is the submission that only made use of *OFF*, we can see that this submission achieved the lowest development and test score for valence. This leads to the assumption that the use of a single feature for predicting emotional valence and arousal tends to come with limited effectiveness and that leveraging other features (e.g., from a different modality) helps in producing better predictions. For Participant 67, we can observe particularly contradictory valence predictions in Fig. 6(A), and where this observation can most likely be attributed to different participant behavior. Indeed, most of the participants freely talk for five minutes without any aid, whereas Participant 67 talks while using written notes. Taking the position of the camera into consideration, the entire face of Participant 67 is hardly visible, and her eyes are focused on the document she has with her. This makes it difficult to extract meaningful facial information (70 keypoints).

For arousal, we can observe two curves with a deviating trend in Fig. 7(A), namely Sub 2 and Sub 3. Specifically, for Sub 3, we can observe that the values obtained are not close to the values obtained by the other submissions, although Sub 3 and the other curves fluctuate similarly (except for Sub 2). For Fig. 7(B), we can observe that all of the curves, including Sub 2, follow a similar trend. Also, looking into the results presented in Table 3, we can again conclude that better arousal predictions can be obtained when making use of multiple features from different modalities. Indeed, when examining Sub 2 and Sub 3 in more detail, we can see that these submissions obtain the lowest arousal development and test scores, with both submissions making use of *OFF* for predicting valence and arousal. In addition, making the comparison to Sub 1, we can observe that the use of *OFF* lowers model effectiveness. Furthermore, given that the curves for Participant 40 and Participant 67 show a similar trend, it is of interest to look at the commonality between these participants in order to achieve a better understanding of the effectiveness of *OFF*. Doing a manual inspection of the audio-visual recordings of these two participants, we notice that the camera angle used is not frontal in nature but rather tilted towards the left. As a result, the face of these participants is not entirely visible in the corresponding recordings, hampering the extraction of all essential features. In this respect, it is also worth noting that in 65 out of the 69 videos available (94.2% of the dataset), the entire face of the participant is visible.

Given the experimental results obtained for late fusion, we can conclude that using multiple features and hyperparameter tuning helps in improving model effectiveness. Additionally, we can conclude that the contradictory results, as present in both Fig. 6(A) and Fig. 7(A), have a significant impact on the model effectiveness when the size of the dataset used is small, even though our experimental results show that the use of a single feature follows a trend that is similar to the trend obtained when combining different features.

We believe that we could still improve model effectiveness during late fusion by applying a more in-depth pre-processing strategy. As an example, and as mentioned in Sect. 2.2, we made use of full-resolution images to extract 25 body key points. Since these full-resolution images contain information unrelated to body key points, the extracted features might contain wrong information. This wrong information may have a negative impact on the model effectiveness obtained during late fusion. As a result, by performing tight cropping so that the images used only contain relevant subjects, we believe the features extracted from these cropped images can help in improving the effectiveness of late fusion.

4 Conclusions and Future Work

In this paper, we explored the simultaneous use of multimodal features, both engineered and learned, to improve human emotion detection for the MuSe-Stress 2021 sub-challenge. Specifically, by applying an LSTM-RNN-based late fusion approach using ResNet-50 and OpenPose features for video, ILPR and eGeMAPS features for audio, and BERT features for text, we achieved a combined CCC value of 0.7595 and 0.3379 for the development set and the test set of MuSe-Stress 2021, respectively. An experimental investigation of the degradation in effectiveness obtained for the test set points to the need for incorporating different strategies to further improve the effectiveness of late fusion, such as data pre-processing and data augmentation.

References

1. Stappen, L., et al.: The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In: Proceedings of the 2nd International on Multimodal Sentiment Analysis Challenge and Workshop. Association for Computing Machinery, New York (2021)
2. Stappen, L., Baird, A., Schumann, L., Schuller, B.: The multimodal sentiment analysis in car reviews (MuSe-car) dataset: collection, insights and improvements. IEEE Trans. Affect. Comput. (2021)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
5. Redmon, J., Farhadi, A.: YOLOV3: an incremental improvement (2018)

6. Baghel, S., Prasanna, S.R.M., Guha, P.: Classification of multi speaker shouted speech and single speaker normal speech. In: TENCON 2017–2017 IEEE Region 10 Conference, pp. 2388–2392. IEEE (2017)
7. Eyben, F., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2015)
8. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Degottex, G.: Glottal source and vocal-tract separation. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI (2010)
10. Rothenberg, M.: Acoustic interaction between the glottal source and the vocal tract. Vocal Fold Physiol. **1**, 305–323 (1981)
11. Loweimi, E., Barker, J., Saz-Torralba, O., Hain, T.: Robust source-filter separation of speech signal in the phase domain. In: Interspeech, pp. 414–418 (2017)
12. Prasanna, S.R.M., Gupta, C.S., Yegnanarayana, B.: Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. **48**(10), 1243–1261 (2006)
13. Baghel, S., Prasanna, S.R.M., Guha, P.: Exploration of excitation source information for shouted and normal speech classification. J. Acoust. Soc. Am. **147**(2), 1250–1261 (2020)
14. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
17. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
18. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018)
19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition, pp. 1–12. British Machine Vision Association (2015)
20. Stappen, L., et al.: MuSe 2020 challenge and workshop: multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: emotional car reviews in-the-wild. In: Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, pp. 35–44 (2020)
21. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: real-time multi-person 2D pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
22. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4645–4653 (2017)
23. Qin, S., Kim, S., Manduchi, R.: Automatic skin and hair masking using fully convolutional networks. In: 2017 IEEE International Conference on Multimedia and Expo (ICME) (2017)
24. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)

25. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. **16**(6), 345–379 (2010)
26. Zhang, Q., Xiao, T., Huang, N., Zhang, D., Han, J.: Revisiting feature fusion for RGB-T salient object detection. IEEE Trans. Circ. Syst. Video Technol. **31**(5), 1804–1818 (2020)