Multimedia Tools and Applications (2021) 80:35871–35885
https://doi.org/10.1007/s11042-020-09842-1

Human emotion recognition based on the weighted integration method using image sequences and acoustic features

Sung-Woo Byun[1] · Seok-Pil Lee[2]

Received: 4 June 2020 / Revised: 27 July 2020 / Accepted: 9 September 2020 /
Published online: 18 September 2020

Abstract
People generally perceive other people's emotions based on speech and facial expressions, so it can be helpful to use speech signals and facial images simultaneously. However, because the characteristics of speech and image data are different, combining the two inputs is still a challenging issue in the area of emotion-recognition research. In this paper, we propose a method to recognize emotions by synchronizing speech signals and image sequences. We design three deep networks. One of the networks is trained using image sequences, which focus on facial expression changes. Facial landmarks are also input to another network to reflect facial motion. The speech signals are first converted to acoustic features, which are used for the input of the other network, synchronizing the image sequence. These three networks are combined using a novel integration method to boost the performance of emotion recognition. A test comparing accuracy is conducted to verify the proposed method. The results demonstrated that the proposed method exhibits more accurate performance than previous studies.

Keywords Emotion recognition · Acoustic feature · Facial expression · Model integration

1 Introduction

Recently, high-performance personal computers have been rapidly popularized with the technological development of information society. Accordingly, the interaction between

✉ Seok-Pil Lee
esprit@smu.ac.kr

Sung-Woo Byun
123234566@naver.com

[1] Graduate School, Department of Computer Science, SangMyung University, Seoul, Republic of Korea

[2] Department of Electronic Engineering, SangMyung University, Seoul, Republic of Korea

humans and computers is actively changing into a bidirectional interface, and a better understanding of human emotions is needed, which could improve human–machine interaction systems [4]. In signal processing, emotion recognition has become an attractive research topic [45]. Therefore, the goal of this human interface is to extract and recognize the emotional state of individuals accurately and to provide personalized media according to a user's emotional state.

Emotion refers to a conscious mental reaction subjectively experienced as strong feeling typically accompanied by physiological and behavioral changes in the body [3]. To recognize a user's emotional state, several studies have applied different forms of input, such as speech, facial expression, video, text, and others [11, 13, 15, 25, 39, 42, 47]. Among the methods using these inputs, facial emotion recognition (FER) has been gaining substantial attention over the past decades. Conventional FER approaches generally have three main steps: 1) detecting a facial region from an input image, 2) extracting facial features, and 3) recognizing emotions. In conventional methods, it is most important to extract appropriate emotional features from the face image. The facial action coding system encodes the movements of specific facial muscles called action units, which reflect distinct momentary changes in facial appearance [8].

In contrast, deep-learning-based FER approaches reduce the dependence between recognition models and preprocessing techniques, such as feature extraction methods, by enabling "end-to-end" learning from outputs to input images. The convolutional neural network (CNN) is the most popular model among several deep-learning models. It convolves input images through many filters and automatically produces a feature map. The feature map is combined with fully connected layers, and the emotional expression is recognized as belonging to a particular class-based output [21]. Recently, various studies have combined facial features and the deep-learning-based model to boost the performance of facial expression recognition [24, 38, 46].

The speech signal is one of the most natural media of human communication. It contains implicit paralinguistic information and linguistic content, including emotion, about the speaker. Several studies have reported that prosodic features, acoustic features, and voice-quality features imply comparatively abundant emotional significance [28]. The most important issue in the speech-emotion recognition system is the effective parallel use of the extraction of proper speech-signal features and an appropriate classification engine. These features include pitch, formant, and energy features [23, 33, 41]. In addition, the mel-frequency cepstrum coefficients (MFCC) feature is representatively used in many studies for speech-emotion recognition [26, 37, 39]. However, because explicit and deterministic mapping between the emotional state and audio features does not exist, speech-based emotion recognition still has a lower rate of recognition than other emotion-recognition methods, such as facial recognition. Therefore, combining appropriate audio features in speech-emotion recognition is critical.

Generally, people recognize the emotions of other people using speech and facial expressions, such as happiness, sadness, anger, and neutrality. According to previous studies, verbal components convey one-third of human communication, and nonverbal components convey two-thirds [19, 29]. Facial expressions represent an example of nonverbal components. In terms of perceptual and cognitive sciences, when a computer infers human emotions, it is natural that using speech signals and the facial images simultaneously can be helpful for accurate and natural recognition. However, because the characteristics of the methods to recognize emotions from speech signals and image sequences are different, combining the two inputs is still a challenging issue in the area of emotion-recognition research.

In this paper, we propose a method to recognize emotions by synchronizing speech signals and image sequences. To do this, we design three deep networks. One of the networks is trained using image sequences, which focuses on facial expression changes. Moreover, facial landmarks are input into another network to reflect facial motion. The speech signals are first converted to acoustic features, which are used for the input of the other network, synchronizing the image sequence. Furthermore, we present a novel method to integrate the models, which performs better than other integrated methods. A test comparing accuracy is conducted to verify the proposed method. The results demonstrated that the proposed method shows better performance than previous studies. Therefore, our main contributions in this paper are summarized as follows:

- Two deep network models recognize emotions from images, and one deep network model recognizes emotions from speech to reflect temporal representations from two kinds of sequential data.
- A method is proposed to learn and classify two different types of data, images and speech, from video data by synchronizing them.
- We present a weighted integration method for these three networks with different characteristics, and performance improvement is achieved in terms of accuracy.

This paper is organized as follows. Section 2 introduces researches on existing emotion recognition. Section 3 explains the proposed emotion recognition method. Section 4 presents the experiment description and results, and then concludes with Section 5.

2 Related work

2.1 Facial emotion recognition

Research on FER has been gaining much attention over the past decades with the rapid development of artificial intelligence techniques. For FER systems, several feature-based methods have been studied. These approaches detect a facial region from an image and extract geometric or appearance features from the region. The geometric features generally include the relationship between facial components. Facial landmark points are representative examples of geometric features [2, 30, 31]. The global facial region features or different types of information on facial regions are extracted as appearance features [20, 36]. The global futures generally include principal component analysis, a local binary pattern histogram, and others. Several of the studies divided the facial region into specific local regions and extracted region specific appearance features [6, 9]. Among these local regions, the important regions are first determined, which results in an improvement in recognition accuracy. In recent decades, with the extensive development of deep-learning algorithms, the CNN and recurrent neural network (RNN) have been applied to the various fields of computer vision. Particularly, the CNN has achieved great results in various studies, such as face recognition, object recognition, and FER [10, 16, 44]. Although the deep-learning-based methods have achieved better results than conventional methods, micro-expressions, temporal variations of expressions, and other issues remain challenging [21].

2.2 Audio emotion recognition

Speech signals are some of the most natural media of human communication, and they have the merit of real-time simple measurement. Speech signals contain linguistic content and implicit paralinguistic information, including emotion, about speakers. In contrast to FER, most speech-emotion recognition methods extract acoustic features because end-to-end learning (i.e., one-dimensional CNNs) cannot extract effective features automatically compared to acoustic features. Therefore, combining appropriate audio features is key. Many studies have demonstrated the correlation between emotional voices and acoustic features [1, 5, 14, 18, 27, 32, 34]. However, because explicit and deterministic mapping between the emotional state and audio features does not exist, speech-based emotion recognition has a lower rate of recognition than other emotion-recognition methods, such as facial recognition. For this reason, finding the optimal feature set is a critical task in speech-emotion recognition.

2.3 Multimodal emotion recognition

Using speech signals and facial images can be helpful for accurate and natural recognition when a computer infers human emotions. To do this, the emotion information must be combined appropriately to various degrees. Most multimodal studies focus on three strategies: feature combination, decision fusion, and model concatenation. To combine multiple inputs, deep-learning technology, which is applied to various fields, can play a key role [7, 22]. To combine the models with different inputs, model concatenation is simple to use. Models inputting different types of data output each encoded tensor. The tensors of each model can be connected using the concatenate function. Yaxiong et al. converted speech signals into mel-spectrogram images for a 2D CNN to accept the image as input. In addition, they input the facial expression image into a 3D CNN. After concatenating the two networks, they employed a deep belief network for the highly nonlinear fusion of multimodal emotion features [28]. Decision fusion aims to process the category yielded by each model and leverage the specific criteria to re-distinguish. To do this, the softmax functions of the different types of networks are fused by calculating the dot product using weights where the summation of the weights is 1. Xusheng et al. proposed a bimodal fusion algorithm to realize speech-emotion recognition, where both facial expressions and speech information are optimally fused. They leveraged the MFCC to convert speech signals into features and combined the CNN and RNN models. They used the weighted-decision fusion method to fuse facial expressions and speech signals [40]. Jung et al. used two types of deep networks—the deep temporal appearance network and the deep temporal geometry network—to reflect not only temporal facial features but also temporal geometry features [17]. To improve the performance of their model, they presented the joint fine-tuning method integrating these two networks with different characteristics by adding the last layers of the fully connected layer of the networks after pre-training the networks. Because these methods mostly use shallow fusion, a more complete fusion model must be designed [28].

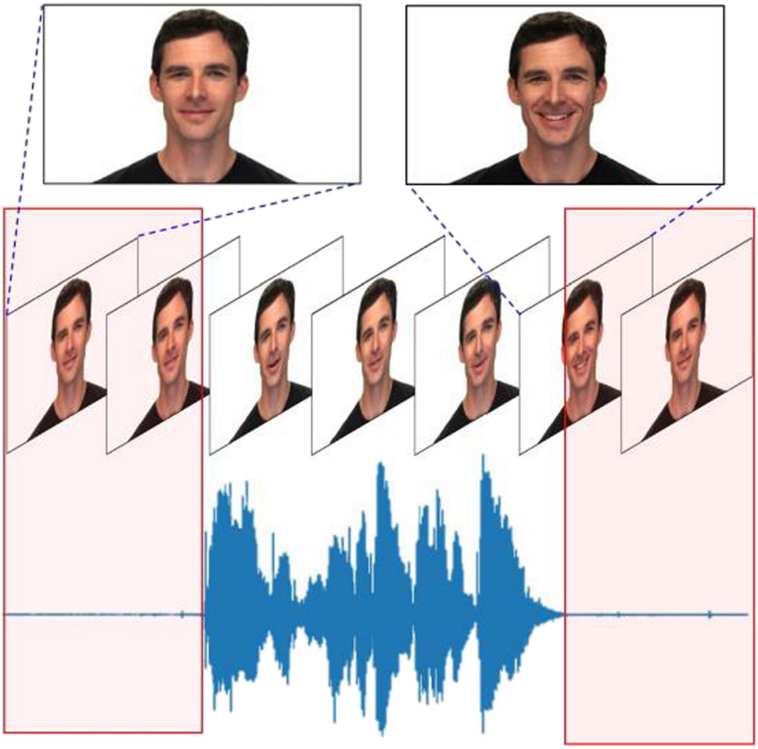

Fig. 1 Audio signal and image sequence from a video; shaded areas indicate when the actor prepares or finishes expressing the emotional state

3 Proposed method

3.1 Preprocessing

When constructing a video emotion database, the actors start and finish expressing emotions according to the instructions of the experimenter. Therefore, as shown in Fig. 1, the database was often divided into three sections—the section for the actor to express emotions, the section to prepare the emotional state, and the section finishing expressing emotions. For this reason, the need to determine whether a given speech signal and image sequence should be classified as the acting section or the silence section arises in many emotion-recognition systems. When nonspeech sections are included in the learning or testing process, they can provide unnecessary information and become an obstacle. For more accurate processes, this section describes removing these nonspeech sections. Because the signal energy value of the speech-signal segment is larger than that of the nonspeech-signal segment, an absolute integral value (IAV) reflecting the energy value was used. The IAV value was computed using Eq. (1):

$$X = \sum_{i=1}^{N} |X(i\Delta t)| \tag{1}$$

where X is the recorded signal,$\Delta t$is the time interval, N is the number of samples, and i is the sample index.

First, the IAV feature vector must be extracted from the interval of the signal. Then, it is imperative to calculate the maximum and minimum values and determine the threshold value with a 10% difference between these two values. An example of determining the threshold is shown in Fig. 2.

The process of selecting the start point for a speech interval includes a point at which the window is larger than the IAV value. If the extracted IAV value was smaller than the IAV threshold, the endpoint was determined. The points were quantized using Eqs. (2) and (3) so that the speech signals and image sequences were synchronized.

$$\textit{Quantization}\ \text{value} = \textit{Sampling rate}/10 \tag{2}$$

$$\begin{cases} \textit{if}\ p = \textit{start}\ \text{point}, & p = \textit{Rounddown}(p/\textit{Quantization value}) \times \textit{Quantization value} \\ \textit{if}\ p = \textit{end}\ \text{point}, & p = \textit{Roundup}(p/\textit{Quantization value}) \times \textit{Quantization value} \end{cases} \tag{3}$$

To map 30 Hz (33.33 ms) of the sampling rate of the image sequence, the window size of the speech signals was 1600 (33.33 ms). Accordingly, the input of an image sequence and speech signal at a point used one image and 1600 speech-signal data, respectively.

3.2 Image-based model

To recognize emotions from a facial image sequence, we used two deep-learning networks. The first network captures temporal changes in appearance by combining the CNN and LSTM models. The proposed CNN and LSTM models are illustrated in Fig. 3.

In general, the length of image sequences varies in every video, but the input length of a deep network is usually fixed. Therefore, the length of the image sequence must be fixed. In this study, we set a time step of the image sequence to ten. The network infers an emotion every 0.3 s. Before inputting an image sequence to the network, all images were converted to grayscale. Then, the faces in the input images were detected, cropped, and rescaled to 64 × 64. The common 2D-CNN layer used still images as input. We combined CNN layers and LSTM layers to deal with image sequences.

The CNN layers of this network used the image sequences as input without sharing weights along the time axis. Thus, the filters played different roles depending on the time. Each image

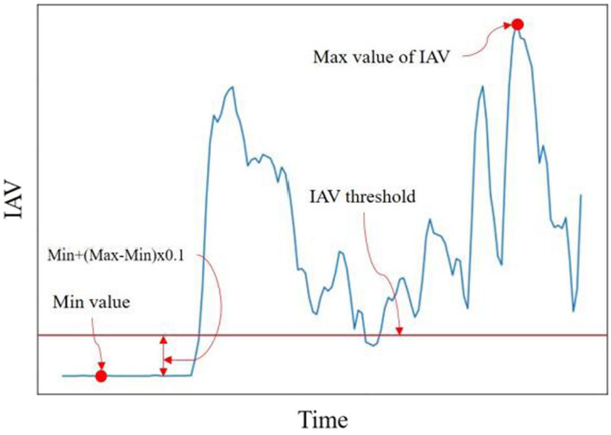

Fig. 2 An example of determining the threshold

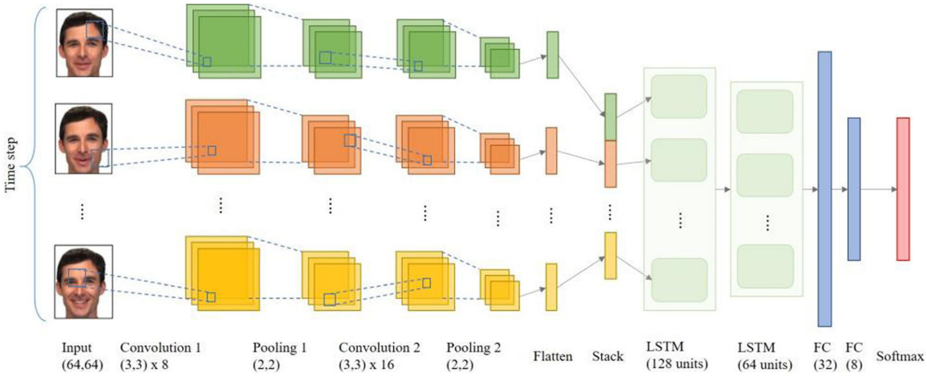

Fig. 3 Structure of the two-dimensional convolutional neural network and long short-term memory (LSTM) model for a facial image sequence

along the time axis was converted to feature maps through each convolutional and pooling layer. After convolving the images, all output passed through rectified linear unit activation functions. The feature maps were stacked in time order so that they were input into the LSTM layers. The output of the LSTM layer was connected with the fully connected layers, and the last layer inferred the probability of each emotion through the softmax function. To train the whole network, the AdaDelta optimizer method was used, and the weight-decay and dropout methods were used for regularization.

The network that input the landmarks was derived from a previous study. Because landmarks generally reflect facial motion, they complement another model to infer facial expression. First, landmarks can be considered to be a 1D vector as follows:

$$X^{(t)} = \left[x_1^{(t)},\ y_1^{(t)},\ x_2^{(t)},\ \mathrm{y}_2^{(\mathrm{t})},\ \cdots,\ x_n^{(t)},\ y_n^{(t)}\right] \tag{4}$$

Where n is the total number of landmark points at frame t, and$X^{(t)}$is a$2 \times n$dimensional vector at t. In addition,$x_k^{(t)}$and$y_k^{(t)}$are coordinates of the k^{th} facial landmark points at frame t. The normalization of the landmark vector was required because each landmark point is a pixel value of the image. The landmark points were normalized based on thexycoordinates of the noise point. The equation is as follows:

$$\widetilde{x}_i^{(t)} = \frac{x_i^{(t)} - x_o^{(t)}}{\sigma_x^{(t)}} \tag{5}$$

where$x_i^{(t)}$is an x-coordinate of the i^{th} facial landmark point at frame t,$x_o^{(t)}$is the x-coordinate of the nose landmark coordinate at frame t, and$\sigma_i^{(t)}$is the standard deviation of the x-coordinates at frame t. This process is also applied to$y_i^{(t)}$. We concatenated the normalized vector along the time step. The vector is used as input to the network, as shown in Fig. 4.

The network receives the normalized vector as input, and the last layer infers the probability of each emotion through the softmax function. The dropout methods are used between each fully connected layer for regularization.

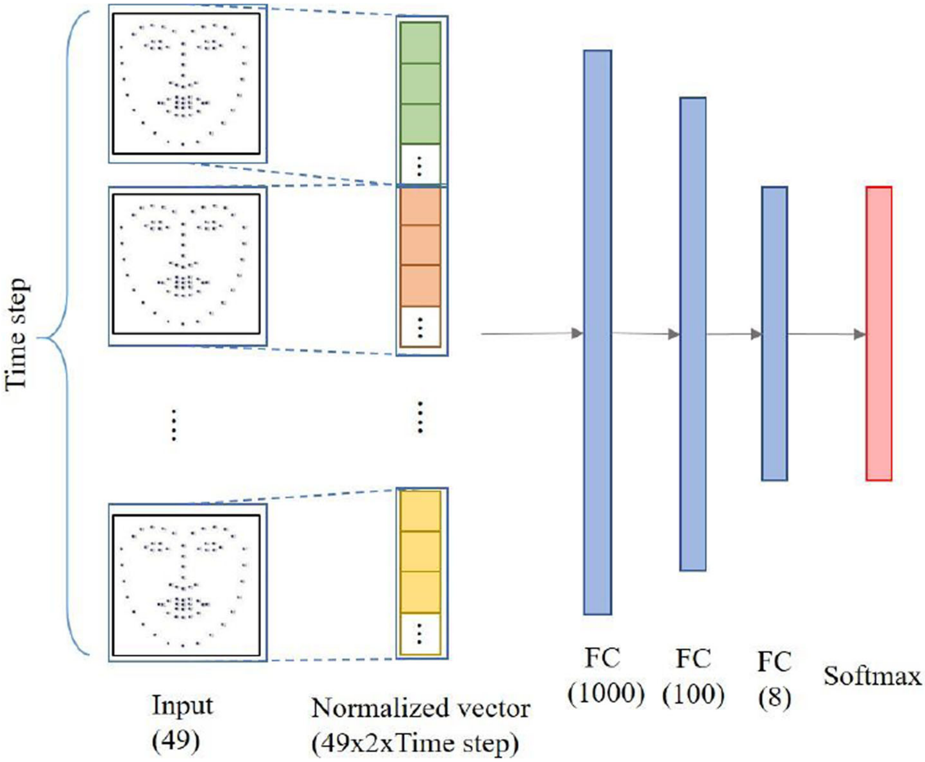

Fig. 4 Structure of the deep neural network for the landmark vector

3.3 Speech-based model

Because verbal components convey one-third of human communication, it is natural that using speech signals and a facial image simultaneously can be helpful for accurate and natural recognition. Therefore, we propose a reasonable feature combination that can improve emotion-recognition performance using an RNN, complementing the FER. In previous emotion-recognition combining speech signals and image sequences, many studies used only the MFCC feature or images converted from a mel-spectrogram [12, 28, 40]. We surveyed acoustic features used for many speech-emotion recognition studies and composed an optimal feature set by analyzing and combining the interconnectivity of each feature. Harmonic features reflecting the harmony of speech are used, which were less used for previous studies. First, we selected specialized features for emotion recognition through individual analysis and found the optimal feature set by recombining features.

In total, 43 features were extracted and are used in this paper:

- 13 MFCCs;
  11 spectral-domain: spectral centroid, spectral bandwidth, 7 spectral contrasts, spectral flatness, and spectral roll-off;
- 12 chroma: 12 dimensional chroma vectors; and.
  7 harmonic features: inharmonicity, 3 tristimuli, harmonic energy, noise energy, and noisiness.

If the range of each attribute value of the learning data is greatly different, the learning will not work efficiently. For example, if the range of a feature vector A is 1 to 1000, the range of another feature vector B is 1 to 10, and the value of A is larger, it seems as if it has a significant effect on the neural network while B seems as if it does not relatively affect the network. Thus, transforming each property value into the same range is necessary before the learning process, and this process is referred to as "feature scaling." In this study, we normalized the features using the standard-score method, which considers the range and variation of the values. The equation of this scaling method is as follows (2).

$$x' = \frac{x-\overline{x}}{\sigma} \tag{6}$$

where $x^{'}$ is a normalized vector, x is an input vector,$\overline{x}$is the average of x, and$\sigma$is the standard deviation of x.

After windowing the speech signals, the signals are converted to acoustic features, and the features are input into the LSTM layers. The output of the LSTM layer is connected with the fully connected layers, and the last layer infers the probability of each emotion through the softmax function. The whole speech-based model is illustrated in Fig. 5. The weight-decay and dropout methods are used for regularization.

3.4 Weighted joint fine-tuning

The previous study [17] proposed a joint fine-tuning method that integrates two networks. After pretraining the networks, the networks were reused. They integrated the two networks by adding the last layers of the fully connected layer of the networks. Then, the linear fully connected networks were retrained, which achieved better results. In this paper, we designed an integration method that weighted each model in the integration process. The last layers were integrated using Eq. (7):

$$W_1 O_I + W_2 O_L + W_3 O_S \tag{7}$$

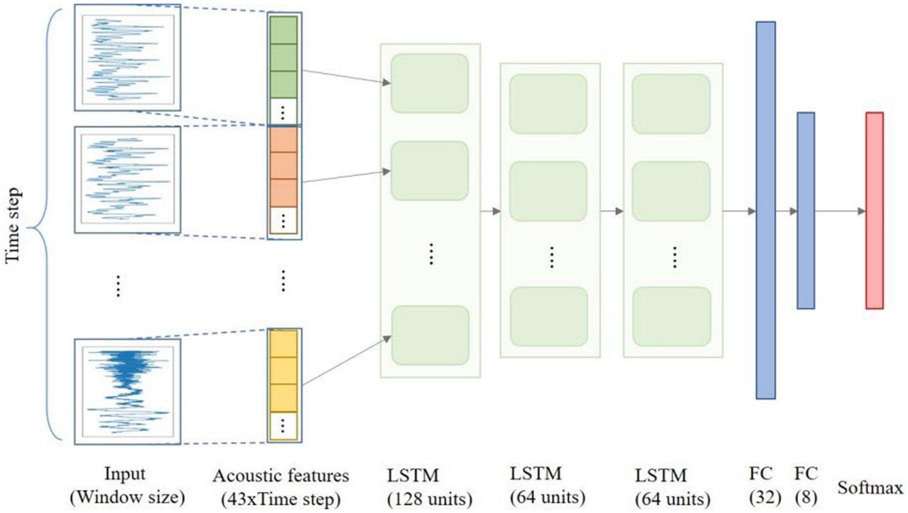

Fig. 5 Structure of the model with acoustic features from speech data

where W_1, W_2, and W_3 are the variables to prioritize the output of each model, and O_I, O_L, and O_S are the output values of the image, landmark, and speech-based model, respectively. Based on the preliminary experiments, we set W_1, W_2, and W_3 to 0.2, 0.2, and 0.6, respectively. Each model was trained using softmax, and pretrained models were integrated using Eq. (7). Finally, the integrated model calculates the probabilities for emotions using another softmax function.

4 Experiment and results

4.1 Ryerson audio-visual database of emotional speech and song dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a video database of emotional speech and songs in North American English, classified into eight emotions as shown in Fig. 6, including neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The database comprises information from 24 professional actors, and each actor has 60 audio-visual (AV) items and 44 song items, for a total of 104 data points. Each recorded production of an actor was available in three modality formats: AV, video only, and audio only. Among these, we used $24 \times 60 \times 3 = 4320$ AV data.

For validation of this database, 247 raters each rated a subset of the 7356 files. For reliability, a further 72 raters provided intra-rater test-retest data. Validation was achieved by asking the raters to label the expressed emotion. In RAVDESS, contrary to traditional validation methods, for facial recognition databases, accuracy, intensity, and genuineness must be verified for emotion measurement of all presented stimuli because orofacial movement, where movements are tied to the lexical content, interacts with movements related to emotional expression. To select the appropriate stimuli, the "goodness" score was imposed. The goodness scores ranged between 0 and 10 and are a weighted sum of the mean accuracy, intensity, and genuineness measures. The equation was defined such that stimuli receiving higher measures of accuracy, intensity, and genuineness were assigned higher goodness scores.

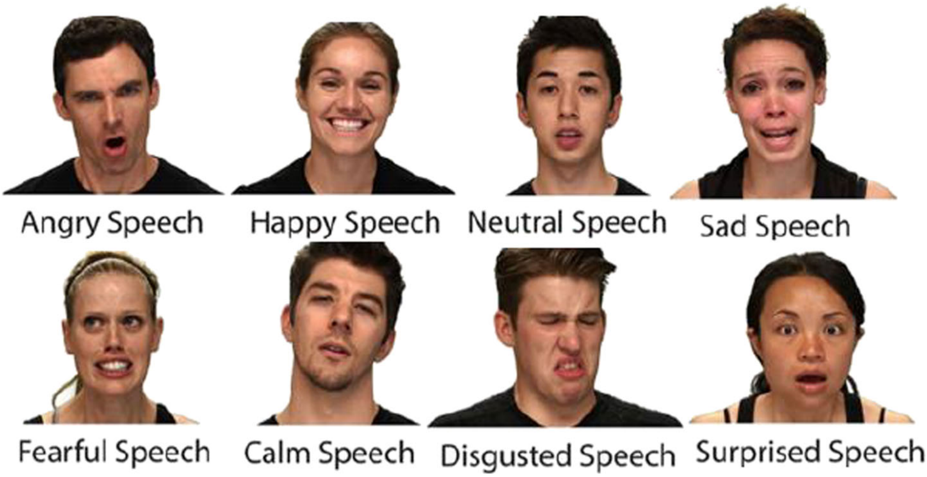

Fig. 6 Examples from the Ryerson Audio-Visual Database of Emotional Speech and Song dataset

4.2 Baselines

This section describes the baseline algorithms for model integration.

4.2.1 Multi-input model

When using different input types for classification, the models should be designed so that each model reflects the characteristics of the data. To combine the models with different inputs, a layer that can connect tensors was used. Models inputting different types of data output each encoded tensor. The tensors of each model can be connected using the concatenate function. Then, the final outputs of each model were integrated by adding the softmax function. Fig. 7 describes the example of the multi-input model.

4.2.2 Feature concatenation

To recognize human emotions by learning both facial and speech data, facial data were converted to a feature map by inputting the data into the CNN. Then, we merged the feature map with features from the speech data, as shown in Eq. (8).

$$x = \{f_1, f_2, \dots f_m, \dots, s_1, s_2 \dots, s_n\} \tag{8}$$

Where f is the feature map from the facial data, and s is the feature of the speech data. Lastly, emotions were classified using the feature vector x as the input of the LSTM model in a time-ordered sequence.

4.2.3 Joint fine-tuning

To incorporate models with different data, first, each model with different data was trained. Only fully connected layers in the pre-softmax classification stage from already trained models were used as new integrated models. The weight values from the already trained models were frozen, and the fully connected layers from each model were retrained. Then, the integrated

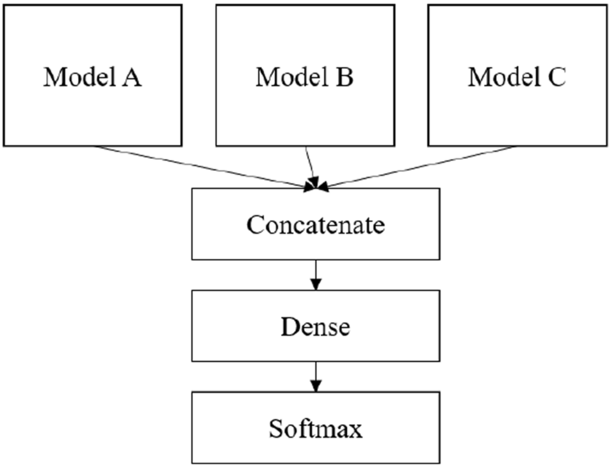

Fig. 7 Examples of the multi-input model

model classified emotions using another softmax. The softmax function for training in each model was used when calculating the loss function, and only the softmax function of the integrated model was used when predicting.

5 Results

As mentioned in Section 4.1, this study used the AV data in the RAVDESS dataset to test performance. The dataset comprised eight emotions, and we used only emotional speech, except the emotional song data. In the RAVDESS, all sequences start and end with a silence section, which was removed through preprocessing (Section 3.1). In addition, image and speech data were separated by synchronizing with each other. The data of a training set were as follows: image data (10, 64, 64), landmark data (980), and speech data (10, 43). To verify the performance of the proposed method, ten-fold cross-validation was performed (Table 1).

When learning the model proposed in this paper using image and speech data with the joint fine-tuning method, the accuracy was 86.06%. When learning the model using the multi-input model and feature combination, the accuracy was 81.93% and 78%, respectively. The model learned using weighted joint fine-tuning demonstrated the greatest result at 87.11%. This also increased the accuracy by about 2.5% compared to the model using only image data (84.69%).

Jung et al. proposed the model recognizing facial expressions using image data, constructing two small deep networks that complement each other [17]. The model proposed by Jung et al. exhibited an accuracy of 85.72% in the RAVDESS dataset. Wang et al. [40], Ma et al.[28], and Hossain et al. [12] proposed models integrating a CNN model input with image data with a 2D CNN model input with speech data by converting the speech signals to mel-spectrogram or spectrogram images. The studies, which converted the speech data to a spectrogram to integrate the image and speech, demonstrated an accuracy of about 75% to 77%. The proposed model integrating image and speech data using acoustic features produced a greater result by about 10% than the other integration methods. The multiple-input model integrating each model using the concatenate function is simple to use, but it may not maximize the ability of the networks. We fine-tuned the softmax functions of the pre-trained networks, considering the characteristic of each input, to maximize the ability of the networks. For this reason, the proposed method can produce more accurate results than the multiple-input model.

Table 1 Comparison results for each study

	Model	Integration Method	Input	Accuracy
1	Proposed	Joint fine-tuning	Image, Speech	86.06%
2	Proposed	Multi-input model	Image, Speech	81.93%
3	Proposed	Feature combination	Image, Speech	78%
4	Proposed	Weighted joint fine-tuning	Image, Speech	**87.11%**
5	Proposed	Weighted joint fine-tuning	Image	84.69%
6	[17]	Joint fine-tuning	Image	82.81%
7	[40]	Multi-input model	Image, Speech	77.66%
8	[28]	Multi-input model	Image, Speech	77.31%
9	[12]	Multi-input model	Image, Speech	75.62%
10	[43]	–	Speech	67.14%
11	[35]	–	Speech	74%

Lastly, most of the previous studies using the RAVDESS dataset used only speech data by converting the speech data to an acoustic feature. They exhibited an accuracy of 64.17% and 74%, respectively. Thus, the proposed model dramatically increased the accuracy (87.11%) by integrating the image and speech data.

6 Conclusions

We presented three networks to reflect the characteristics of each input data. One of the networks was trained using image sequences, which focus on facial expression changes. In addition, facial landmarks were input into another network to reflect facial motion. The other network used acoustic features from speech data as input. These three networks were combined using a novel integration method to boost the performance of emotion recognition. To investigate the performance of our model, we tested the recognition accuracy with previous studies on the RAVDESS dataset. According to the results, our model achieved the best recognition rate against facial and speech-based studies. Furthermore, we demonstrated that our weighted joint fine-tuning method exhibited better performance than other methods.

References

1. Bjorn S, Stefan S, Anton B, Alessandro V, Klaus S, Fabien R, Mohamed C, Felix W, Florian E, Erik M, Marcello M, Hugues S, Anna P, Fabio V, Samuel K (2013) Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism
2. Deepak G, Joonwhoan L (2013) Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 13:7714–7734. https://doi.org/10.3390/s130607714
3. Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos J, Delahoz EJ, Contreras-Ortiz S (2020) A machine learning model for emotion recognition from physiological signals. Biomed Signal Proces 55:101646
4. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 44:572–587. https://doi.org/10.1016/j.patcog.2010.09.020
5. Eyben F, Scherer KR, Schuller BW et al (2016) The Geneva minimalistic acoustic parameter set (geMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7:190–202. https://doi.org/10.1109/TAFFC.2015.2457417
6. Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. Multimed Tools Appl 76:7803–7821. https://doi.org/10.1007/s11042-016-3418-y
7. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. https://www.deeplearningbook.org. Accessed 1 Mar 2020
8. Hamm J, Kohler CG, Gur RC, Verma R (2011) Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. J Neurosci Methods 200:237–256
9. Happy SL, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In Proc 4th Int Conf Intell Human Comput Interact 27–29:1–5

10. Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Conf Comput Vision Pattern Recognit Workshops (CVPRW). https://doi.org/10.1109/CVPRW.2017.282
11. He J, Li D, Bo S, Yu L (2019) Facial action unit detection with multilayer fused multi-task and multi-label deep learning network. KSII Trans Internet Inf Syst 7:5546–5559. https://doi.org/10.3837/tiis.2019.11.015
12. Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio–visual emotional big data. Inf Fusion 49:69–78. https://doi.org/10.1016/j.inffus.2018.09.008
13. Hutto CJ, Eric G (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. AAAI Publications, Eighth Int AAAI Conf Weblogs Soc Media
14. Iliou T, Anagnostopoulos C-N (2009) Statistical evaluation of speech features for emotion recognition. In: Digital telecommunications ICDT'09 4th Int Conf IEEE 121–126
15. Jia X, Li W, Wang Y, Hong S, Su X (2020) An action unit co-occurrence constraint 3DCNN based action unit recognition approach. KSII Trans Internet Inf Syst 14:924–942. https://doi.org/10.3837/tiis.2020.03.001
16. Joseph R, Santosh D, Ross G, Ali F (2015) You Only Look Once: Unified, Real-Time Object Detection arXiv preprint arXiv:1506.02640
17. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. 2015 IEEE Int Conf Comput Vision (ICCV) https://doi.org/10.1109/ICCV.2015.341
18. Kao YH, Lee LS (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: InterSpeech
19. Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expression database—A validated database of emotional and conversational facial expressions. PLoS One 7:e32321.
20. Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recogn Lett 34:1159–1168. https://doi.org/10.1016/j.patrec.2013.03.022
21. Ko BC (2018) A brief review of facial emotion recognition based on visual information. Sensors 18. https://doi.org/10.3390/s18020401
22. LeCun Y, Bengio Y, Hinton G (2015) Deep learning, Nature 521. https://doi.org/10.1038/nature14539
23. Lee C, Lui S, So C (2014) Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition. J Acoust Soc Am 135:2422. https://doi.org/10.1121/1.4878044
24. Li S, Deng W (2020) Deep facial expression recognition: A survey. IEEE Trans Affective Comp (Early Access). https://doi.org/10.1109/TAFFC.2020.2981446
25. Liu M, Li S, Shan S, Wang R, and Chen X (2014) Deeply learning deformable facial action parts model for dynamic expression analysis. 2014 Asian Conference on Computer Vision (ACCV) 143–157. https://doi.org/10.1007/978-3-319-16817-3_10
26. Lotfian R, Busso C (2019) Curriculum learning for speech emotion recognition from crowdsourced labels. IEEE/ACM Trans Audio, Speech Lang Processing 4. https://doi.org/10.1109/TASLP.2019.2898816
27. Luengo I, Navas E, Hernáez I, Sánchez J (2005) Automatic emotion recognition using prosodic parameters. In: Interspeech, 493–496
28. Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A (2019) Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Inf Fusion 46:184–192. https://doi.org/10.1016/j.inffus.2018.06.003
29. Mehrabian A (1968) Communication without words. Psychol Today 2:53–56
30. Mira J, ByoungChul K, JaeYeal N (2016) Facial landmark detection based on an ensemble of local weighted regressors during real driving situation. Int Conf Pattern Recognit 1–6.
31. Mira J, ByoungChul K, Sooyeong K, JaeYeal N (2018) Driver facial landmark detection in real driving situations. IEEE Trans Circuits Syst Video Technol 28:2753–2767. https://doi.org/10.1109/TCSVT.2017.2769096
32. Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. Int J Speech Technol 16(2):143–160
33. Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. Speech Comm 40: 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5. https://www.sciencedirect.com/science/article/pii/S0167639302000845. Accessed 1 Mar 2020
34. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Comm 53(9–10):1062–1087. https://doi.org/10.1016/j.specom.2011.01.011
35. Shaqr FA, Duwairi R, Al-Ayyou M (2019) Recognizing emotion from speech based on age and gender using hierarchical models. Procedia Comput Sci 151:37–44. https://doi.org/10.1016/j.procs.2019.04.009
36. Siddiqi MH, Ali R, Khan AM, Park YT, Lee S (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans Image Proc 24:1386–1398. https://doi.org/10.1109/TIP.2015.2405346

37. Song P, Zheng W (2018) Feature selection based transfer subspace learning for speech emotion recognition. IEEE Trans Affective Comput (Early Access) https://doi.org/10.1109/TAFFC.2018.2800046
38. Sun N, Qi L, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn Lett 119:49–61. https://doi.org/10.1016/j.patrec.2017.10.022
39. Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: A review. Int J Speech Technol 21:93–120. https://doi.org/10.1007/s10772-018-9491-z
40. Wang X, Chen X, Cao C (2020) Human emotion recognition by optimally fusing facial expression and speech feature. Signal Process Image Commun https://doi.org/10.1016/j.image.2020.115831
41. Wu CH, Yeh JF, Chuang ZJ (2009) Emotion perception and recognition from speech, Affective Inf Processing 93–110. https://doi.org/10.1007/978-1-84800-306-4_6.
42. Xiong X and Fernando DlT (2013) Supervised descent method and its applications to face alignment. 2013 IEEE Conf Comput Vision and Pattern Recognit (CVPR) https://doi.org/10.1109/CVPR.2013.75
43. Zamil AAA, Hasan S, Baki SJ, Adam J, Zaman I (2019) Emotion detection from speech signals using voting mechanism on classified frames. 2019 Int Conf Robotics, Electr Signal Processing Technol (ICREST) https://doi.org/10.1109/ICREST.2019.8644168
44. Zhang H, Huang B, Tian G (2020) Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. Pattern Recogn Lett 131:128–134. https://doi.org/10.1016/j.patrec.2019.12.013
45. Zhang S, Zhang S, Huang T, Gao W (2008) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans Multimed 20:1576–1590. https://doi.org/10.1109/TMM.2017.2766843
46. Zhang T, Zheng W, Cui Z, Zong Y, Yan J, Yan K (2016) A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans Multimed 18:2528–2536. https://doi.org/10.1109/TMM.2016.2598092
47. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed Signal Processing Control 47:312–323. https://doi.org/10.1016/j.bspc.2018.08.035

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.