sensors

MDPI

Article

Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features

Dilnoza Mamieva [1], Akmalbek Bobomirzaevich Abdusalomov [1], Alpamis Kutlimuratov [2], Bahodir Muminov [3] and Taeg Keun Whangbo [1,*]

[1] Department of Computer Engineering, Gachon University, Seongnam-si 13120, Republic of Korea; mamiyeva.dilnoza@gmail.com (D.M.)
[2] Department of AI. Software, Gachon University, Seongnam-si 13120, Republic of Korea
[3] Department of Artificial Intelligence, Tashkent State University of Economics, Tashkent 100066, Uzbekistan
[*] Correspondence: tkwhangbo@gachon.ac.kr

Abstract: Methods for detecting emotions that employ many modalities at the same time have been found to be more accurate and resilient than those that rely on a single sense. This is due to the fact that sentiments may be conveyed in a wide range of modalities, each of which offers a different and complementary window into the thoughts and emotions of the speaker. In this way, a more complete picture of a person's emotional state may emerge through the fusion and analysis of data from several modalities. The research suggests a new attention-based approach to multimodal emotion recognition. This technique integrates facial and speech features that have been extracted by independent encoders in order to pick the aspects that are the most informative. It increases the system's accuracy by processing speech and facial features of various sizes and focuses on the most useful bits of input. A more comprehensive representation of facial expressions is extracted by the use of both low- and high-level facial features. These modalities are combined using a fusion network to create a multimodal feature vector which is then fed to a classification layer for emotion recognition. The developed system is evaluated on two datasets, IEMOCAP and CMU-MOSEI, and shows superior performance compared to existing models, achieving a weighted accuracy WA of 74.6% and an F1 score of 66.1% on the IEMOCAP dataset and a WA of 80.7% and F1 score of 73.7% on the CMU-MOSEI dataset.

Keywords: CNN; multimodal emotion recognition; facial feature; speech feature; attention mechanism

check for updates

Citation: Mamieva, D.; Abdusalomov, A.B.; Kutlimuratov, A.; Muminov, B.; Whangbo, T.K. Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. *Sensors* **2023**, *23*, 5475. https://doi.org/10.3390/s23125475

Academic Editor: Antonio Fernández-Caballero

Received: 24 April 2023
Revised: 25 May 2023
Accepted: 8 June 2023
Published: 9 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Emotions are multifaceted psychological phenomena that permeate interpersonal interactions and have far-reaching effects on people's actions [1]. Recognizing and understanding others' emotions while communicating is crucial for meaningful conversations. Emotion recognition based on a single modality, such as facial expressions or voice, is difficult and frequently inaccurate [2,3]. Multimodal emotion recognition [4,5] has been developed to address this restriction. The goal of multimodal emotion recognition is to enhance the reliability of emotion identification systems by including data from many modalities, such as facial expressions [6,7], spoken words [8–10], and text [11,12]. A person's emotional state can be captured more accurately by combining many senses. This has sparked a renewed push in recent years to create multimodal systems to identify human emotions.

Facial and vocal expressions are two key components in the process of identifying emotions. Speech transmits information about the tone, prosody, and substance of communication, whereas facial expressions provide visual indications of the emotional state of a person. Using facial expressions alone provides various obstacles and constraints in spite of the fact that they are an important modality for identifying emotions. Other factors, such as lighting, occlusion, and head movement, may also have an impact on face emotions. Furthermore, facial expressions are not always visible since people may try to disguise their emotions, either intentionally or unintentionally. As a result, some phrases may be disregarded. Unfortunately, the accuracy of speech-based emotion identification systems may be impacted by the fact that people have varying accents, dialects, speech speeds, and pronunciations. It is also difficult to capture all important elements for reliable emotion detection since emotions may be expressed via several parts of speech, including prosody, intonation, and lexical content. Therefore, it may be impossible to reliably identify complex emotions, such as mixed emotions, including many states of feeling, based just on facial and vocal expressions. To give complete and accurate insight into a person's emotional state, it is often important to use both characteristics together.

In the case of multimodal emotion recognition from facial and speech features, deep neural networks [13–15] have been employed to extract relevant features from each modality. Networks learn to extract features relevant to emotion classification. For example, in facial expression analysis, deep neural networks can learn to extract features such as facial landmarks [16], head poses, and eye gaze directions. Similarly, deep neural networks may learn to derive characteristics such as pitch, amplitude, and spectral information from speech. However, the complexity of the system in multimodal emotion recognition can be computationally challenging, requiring significant resources and time to train and execute deep neural networks. In addition, overfitting may occur in deep neural networks, which can result in a poor data performance that has never been observed before. The complex and ever-changing characteristics of emotions may also make it difficult for deep neural networks to recognize them reliably under real-world circumstances. By narrowing their focus on the most relevant aspects of the input data, attention mechanisms [17–19] have been proven to boost the effectiveness of deep neural networks. This is crucial because not all sensory modalities contribute equally to emotion categorization. Improved emotion detection accuracy may result from the system's use of the attention mechanism [20–22] to zero down on and prioritize the most important modalities. Facial expressions and vocal characteristics may have various feature vector dimensions when multimodal emotion recognition is addressed. Attention mechanisms can be used to equalize the dimensions of feature vectors before feeding them into a neural network, which can lead to more accurate emotion recognition.

In this work, we applied an attention mechanism to improve multimodal emotion identification through vocal and facial characteristics. The suggested emotion recognition system in this research has two primary components: speech feature encoder and facial feature encoder. The facial feature encoder extracts high and low facial features from images using a convolutional neural network (CNN), whereas the speech feature encoder uses the Mel-frequency cepstral coefficients (MFCCs) via CNN to ensure a stable training process for the spectral and time information and waveform features to avoid losing important information when dealing with speech data of varying lengths. After extracting the features from the two modalities, we used an attention mechanism to select the most important features for each modality. The attention mechanism takes the feature vectors from the facial and speech modalities as input and computes an attention weight for each feature vector. The attention weight reflects the importance of each feature vector in the overall emotion recognition task. The attention weights for facial and speech features are combined using a fusion network to create a multimodal feature vector. This multimodal feature vector contains the most important features from both modalities and is fed to a classification layer for final emotion recognition. Our experimental results demonstrate the reliability of the suggested system on the IEMOCAP and CMU-MOSEI datasets. The proposed model has the potential to be used in various applications, including affective computing, human–robot interaction, and mental health diagnosis.

This work contributed to the area of multimodal emotion recognition in numerous ways:

- This study suggests a novel way for identifying multimodal emotions by bringing together facial and verbal clues with an attention mechanism. This method addresses the shortcomings of unimodal systems and enhances the accuracy of emotion recognition by using valuable data from both modalities.
- Time and spectral information were used to address the challenges posed by varying the length of speech data. This allows the model to focus on the most informative parts of the speech data, thereby reducing the loss of important information.
- Facial expression modalities involve generating low- and high-level facial features using a pretrained CNN model. Low-level features capture the local facial details, whereas high-level features capture the global facial expressions. The use of both low- and high-level features enhances the accuracy of emotion recognition systems because it provides a more comprehensive representation of facial expressions.
- This study improves the generalization of the multimodal emotion recognition system by reducing the overfitting problem.
- Finally, the attention mechanism is effectively utilized to focus on the most informative parts of the input data and handle speech and image features of different sizes.

This article's remaining sections are structured as follows: In Section 2, recent research related to multimodal emotion recognition, including speech and facial expression modalities, and other DL methodologies that integrate attention mechanisms are discussed. Sections 3 and 4 present and explain, in detail, the workflow of the proposed multimodal emotion recognition system, and the empirical results of the proposed model, including a performance comparison with benchmark models. Section 5 concludes the paper by summarizing the contributions of the proposed model and discussing potential future directions for improving the proposed system. Finally, a list of recent referenced studies is provided.

2. Related Works

In recent years, there has been growing interest in multimodal emotion recognition, driven by advances in deep learning and signal processing techniques. Researchers have proposed and tested various methods [23–26] to achieve high accuracy in multimodal emotion recognition, and this field has seen significant progress in terms of both accuracy and real-world applications. Human emotions are subjective and may be influenced by elements such as cultural background, personality, and situational context, making multimodal emotion detection challenging. As a result, developing a single model that can reliably anticipate emotions for all circumstances and people is challenging. A number of approaches [27,28] for addressing this issue have been suggested, including customizing multimodal emotion detection algorithms to particular circumstances and users. Inspired by PathNet's success in multi-task learning, the research [27] presents a meta-transfer learning strategy for emotion identification, testing its efficacy in transferring emotional information across visual datasets and considering its potential for voice signals. Moreover, in terms of model generalization, the authors of the paper [5] suggest a framework for facial emotion recognition that involves a pre-trained spatial transformer network on saliency maps and facial images, followed by a bi-LSTM that incorporates an attention mechanism. This study used only one dataset, which may limit the generalizability of the proposed dynamic fusion model to other datasets.

Dealing with the diversity of human expressions is a significant obstacle in recognizing facial expressions, making it challenging to create a universal model that can precisely identify emotions in all circumstances and individuals. Thus, this article [29] proposes a multi-modal method for extracting emotion features from facial expression images by combining low-level empirical features with high-level self-learning features obtained through CNNs. The former is extracted from the 2D coordinates of facial key-points, whereas the latter is obtained from the CNNs. Although several methods [30,31] have been proposed for facial expression recognition, there is still room for improvement in terms of accuracy and generalization.

Moreover, recognizing emotions from speech signals has shown great potential with the use of deep learning-based methods, particularly with the implementation of CNNs and RNNs [32–34]. The article [35] proposes an approach for improving speech emotion recognition which enhances speech features by selecting specific subsets of the feature set, applying principal component analysis to these subsets, fusing the resulting features horizontally, analyzing the feature set using t-SNE, and then using the features for emotion recognition.

Most emotion recognition methods use only one of these sources. Emotion recognition models that rely on a single source of information may be easier to implement; however, they are more likely to contain errors and inaccuracies owing to the limited scope of the input data.

Alternatively, multimodal models have the potential to mitigate the deficiencies of relying on a single information source, thereby improving the accuracy and robustness of emotion recognition. Information overload, which may occur when multiple sources of information are combined in multimodal models, can be addressed using an attention mechanism. There has been different research [20–22,36–38] on the attention mechanisms used in multimodal emotion recognition. The authors of [20] proposed a fully end-to-end model for multimodal emotion recognition that jointly optimizes both phases by taking raw input data, conducting data restructuring, and learning features automatically through end-to-end training, resulting in improved task-specific feature optimization and the elimination of manual feature extraction. However, it introduces higher computational overhead and potential overfitting. To address these concerns, they integrated a sparse cross-modal attention mechanism and sparse convolutional neural network (CNN) to select relevant features, reduce redundancy, and mitigate noise in the input data. In the article [21], video emotion recognition approach emphasizes representation learning and enhances the encoders of audio and visual modalities using global semantic information from text. Additionally, the approach reinforces the audio and visual feature sequences by integrating complementary information from each modality and employing attentive decision fusion to obtain the ultimate emotion prediction.

In addition, this method [39] eliminates the need to detect and follow facial landmarks in emotion recognition systems based on video, which is a common source of errors, thereby improving the resilience of video processing. In addition, the audio component of the system employs Gaussian mixture models (GMMs) specific to utterances derived from a Universal Background Model (UBM) using MAP estimation.

Overall, attention-based multimodal emotion recognition models provide a solution to overcome the challenge of information overload and adapt to varying input uncertainties, leading to improved accuracy and robustness in emotion recognition. As such, these models represent a significant advancement in the field of affective computing, with potential applications in various domains, including healthcare, education, and entertainment.

3. The Proposed System

Our research presents a new method for emotion detection by combining vocal and facial cues with an attention mechanism. Multiple elements make up the modalities, and they all contribute to the emotion prediction process in their own way. Speech modality is modeled using a convolutional neural network (CNN), whereas image modality is modeled using the ResNet model. An attention mechanism is employed to weigh the importance of specific features in emotion recognition. A detailed illustration of the modeling process is presented in Figure 1, which shows the flow of the various components involved in the model.

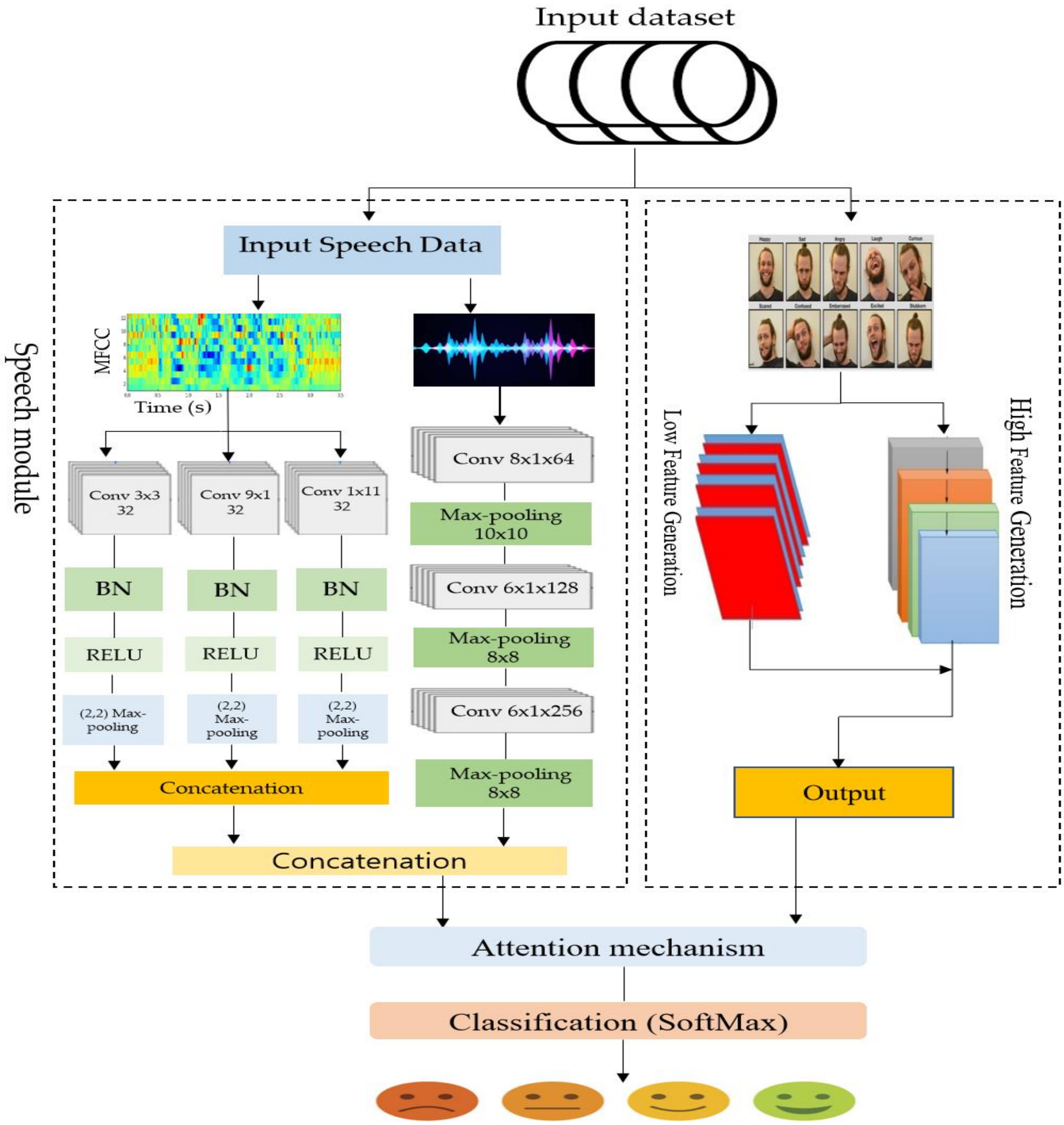


Figure 1. The workflow of the developed emotion recognition system.

3.1. Facial Module

The facial feature extraction module is the first component of the proposed emotion recognition that is responsible for extracting features from facial images. It is designed [40] to encode visual information in the form of static images, such as facial expression, eye gaze, and head pose, into a low-dimensional feature vector that can be further processed by the downstream model for emotion recognition. The facial feature extraction module employs ResNet [41] to capture local and global patterns in facial images and generate high-level abstract features that are discriminative for different emotions.

The high feature generation pyramid (HFGP) is used for feature extraction and representation from images. It combines shallow and deep features to obtain a more comprehensive representation of the input image. In our facial module (Figure 2), ResNet provides multilevel semantic information for feature maps through its conv4 and conv5 layers. By combining these shallow and deep features, the HFGP can create a more comprehensive representation of the image, capturing both low- and high-level details.

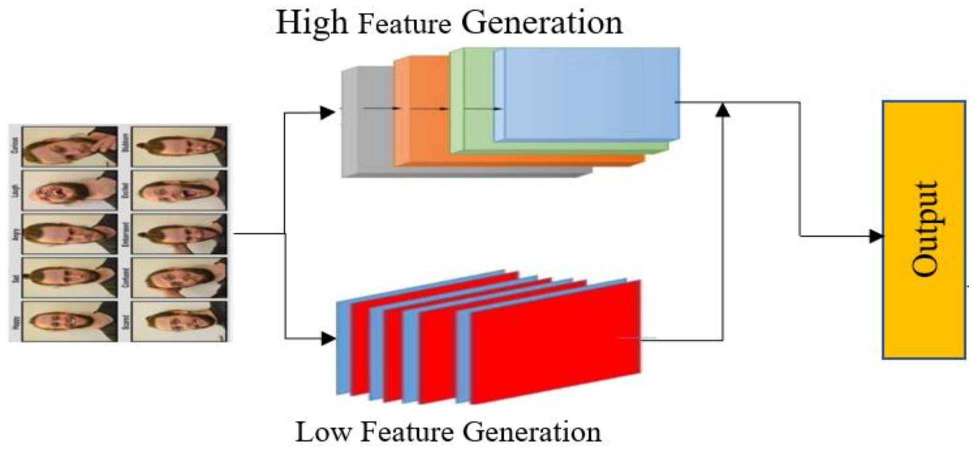


Figure 2. Facial feature extraction model.

The precision of picture categorization may benefit from this.

The low feature generation pyramid (LFGP) and convolution layers are stacked alternately as the second phase in the feature creation process. The LFGP is responsible for generating low-level feature maps with a scale different from that of the HFGP, whereas the convolution layers combine the main features and the large output feature map of the preceding pyramid-based layers. This helps further refine and extract relevant features from the input image data. The added feature maps from the low feature generation pyramid are fed into the next convolutional layer, which analyzes and learns the properties of the feature maps from the pyramid layers and considers them as the foundational features of F_{fo}.

$$\left[f_1^l, f_2^l, f_3^l, \ldots f_j^l\right] = \left\{\begin{array}{cc} T_l\left(F_{fo}\right), & l = 1 \\ T_l(P\left(F_{fo,f_j^{l-1}}\right), & l = 2, \ldots L \end{array}\right\} \tag{1}$$

Equation (1) indicates that the output multiscale features are obtained by combining the features from the base layer F_{fo} with those from each scale (f_j^l) within each LFGP layer. These combined features are then processed by the HFGP layer (T_l) and overall HFGP processing (P) to produce the final output features. The HFGP method is crucial for creating the final multi-level feature pyramid by fusing features from one level in the network. The 1 × 1 convolution layers utilizing the channels of the input features are used for the compression and coupling operations, which combine the feature maps. The HFGP is particularly effective in detecting small objects because it can rescale deep functions to the same scale as the coupling operation and extract high-decision prototypes for better functional extraction.

Low Feature Generation Pyramid (LFGP)

The pyramid network comprises a series of convolution layers with a 2-stride and 3 × 3 kernel. The output of these layers is used as the input for the subsequent convolution layers to generate feature maps. The final layer at each level is selected using the lower convolution layer in the HFGP backbone. To maintain feature smoothness and enhance the learning ability, 1 × 1 convolution layers are added after up-sampling, and a detailed explanation is provided on how the creative sum worked within the top convolution layer network. The multiscale features of the present level are generated by combining the outputs of each convolution layer in both HFGP and LFGP.

3.2. Speech Module

The speech module encoder is a key component of the proposed emotion recognition system, which extracts relevant features from the input speech data. To build the speech module, we used a component from our previous study [42] as the basis. The speech module encoder consists of two branches: one for processing the MFCCs and the other for processing the waveform. The MFCC and waveform branches comprise several layers of convolutional units that learn to extract relevant features from the MFCCs and raw waveform signals. The outputs of both branches are then concatenated, and the concatenated vector as a speech module feature is passed with facial module features through a self-attention mechanism that learns to weigh the contributions of each feature module based on its importance in the emotion prediction task.

3.2.1. MFCC Feature Extractor

CNN blocks of different sizes are designed to facilitate the consolidation of the training stage for spectral and time information. Furthermore, it has been shown empirically [43] that an improvement in the accuracy of predictions is correlated with an increase in the size of the effective area of the CNN. An augmented receptive field size leads to a surge in the number of model parameters, ultimately causing model overfitting, as stated in [44]. Considering the aforementioned objectives, the construction of the model components involves the utilization of three CNNs that are parallelly situated with varying filter sizes. The purpose is to extract different feature maps from the MFCC. The resulting features are then concatenated to form the final output. To compute the MFCC for the input speech data, normalization is first performed, followed by windowing to obtain 64-ms divided frames. Subsequently, Fourier transform is applied to each frame to obtain the frequency components. The next step involves computing an initial set of 40 coefficients for each MFCC frame, using an inverse cosine transform, which is then utilized to train the CNN. The following techniques are used to build the CNN blocks:

- expanding the CNN's depth by incorporating additional layers
- implementing average pooling or larger stride
- making use of expanded convolutions
- employing a separate convolution on each channel of an input

By incorporating additional layers and enlarging the size of the convolution kernel (as depicted in Figure 3), a deep CNN is formed to expand the receptive field. Overfitting on complex dimensions is addressed by computing the receptive field individually per dimension, as indicated in reference [43]. The CNN blocks are utilized with different convolution kernel sizes (3×3, 9×1, and 1×11) to capture spectral and time information, resulting in reduced computational complexity and fewer model parameters compared to a single CNN block with a similar effective area size. In the interpretation time, feature-wise actions are performed by batch normalization (BN) where the CNN's effective area remains constant, and the BN parameters are generated from the effective area and each layer's activations of the raw speech input. The "spots" of a convolutional kernel are obtained through dilations, resulting in the elimination of kernel weights in the spatial neighboring of samples despite their unchanged number. The convolution process for calculating is performed on the samples by the kernel with a striding factor of "α" when it is diluted. According to reference [43], layers that use dilations make use of an expanded spatial length "$\alpha(k - 1) + 1$" as the kernel spatial length increases to that value due to dilation. In addition, convolutions can be distinguished based on their channel or spatial dimensions, and these distinguished convolutions have the same receptive field characteristics as their parallel equivalents. To compute the effective area, a kernel size of "3" is used in the 3×3 depth-wise convolution, and the resulting encoded MFCC features from each CNN block are merged.

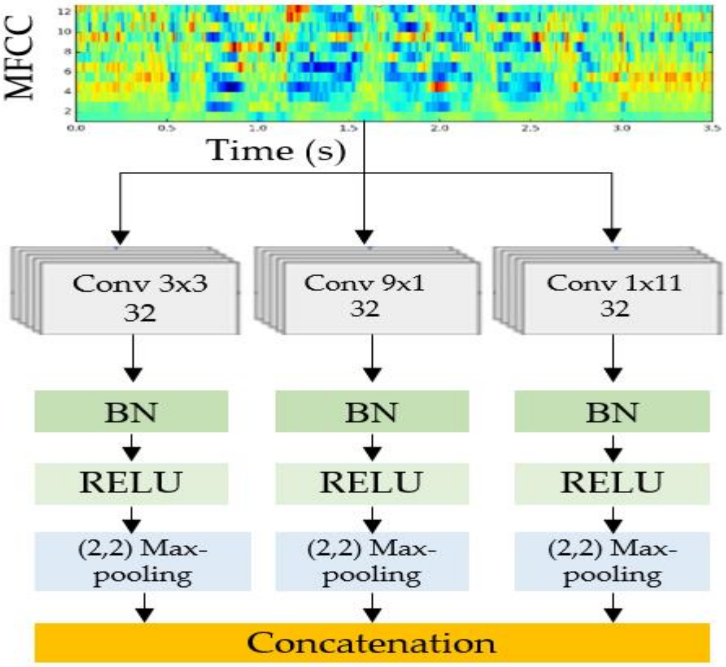


Figure 3. MFCC feature extractor.

3.2.2. Waveform Feature Extractor

To maintain constancy during the training phase of the proposed model and reinforce its generalization, an attempt is made to incorporate paralinguistic information, as it is believed that the combination of various crucial features, as shown in several developed models [45,46], might lead to better performance. The waveform feature extractor (WFE) comprises a triplet of successive convolutional layers that perform a computation process based on Equation (2), where the input waveform data undergoes the kernel function denoted by f(x).

$$(f * x)(T) = \sum_{s=-t}^{t} f(T) \times (T - s) \tag{2}$$

After obtaining unit variance and zero mean, the input waveform data are partitioned into 20-s intervals and fed to the convolutional layer, which is followed by a max pooling operation that reduces the dimensionality. A key factor in selecting an appropriate pooling size is the size of the convolution kernel. A straightforward empirical approach is adopted, as shown in Equation (3).

$$M = \frac{L - 1}{L + X - 1} \tag{3}$$

Equation (3) is used to determine an appropriate pooling size based on the convolution kernel size, with “X” denoting pooling size, “L” denoting kernel size, and “M” denoting overlap rate, typically assumed to be approximately 0.5 for hand-crafted features. The complete information is considered by strides, whereas max pooling only concentrates on the most important information and eliminates the irrelevant data; therefore, to prevent acquiring the same characteristics for succeeding frames, the value of “M” should be less than 0.5. As a result, when constructing the WFE architecture, the value of “M” is considered and established as “0.41” for the first convolutional layer and “0.38” for the second and third convolutional layers, as seen in Figure 4.

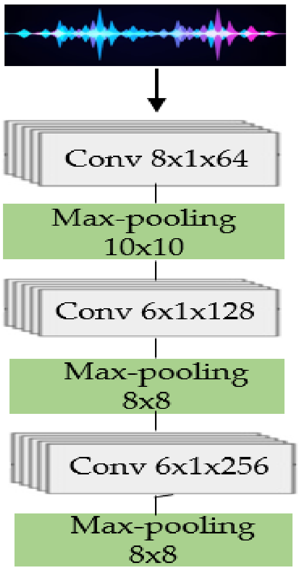


Figure 4. Waveform FE components.

The WFE structure involves different strides and kernel sizes for the convolution and max pooling. The initial convolutional layer utilizes 64 filters in the temporal domain with an "L" of 8, followed by a down-sampling technique called max pooling, reducing the frame rate with an "X" of 10. The second convolutional layer has a channel size of 128 and an "L" of 6, while the second max pooling layer has a size of 8 and an "M" factor below 0.5. The third convolutional layer has a filter size of 256 and an "L" of 6. A max pooling layer with a dimension of 8 is deployed over the time domain as the final step.

3.3. Attention Mechanism

The performance of deep neural networks can be enhanced by an attention mechanism that selectively concentrates on the most important aspects of an object for classification. This enables the attention mechanism to improve the accuracy of the original model. It has been shown that an attention mechanism is useful in natural language processing tasks, such as sentiment analysis [47], where it is used to determine which words and phrases are most important in a sentence to forecast the author's intent. This has led to its use in other areas, such as multimodal emotion identification, where it has been demonstrated to enhance model performance by focusing attention only on where it will perform the best. The contributions of different modalities may vary in terms of their relevance to sentiment-classification tasks. Some modalities may be more informative and contribute more significantly to a task than others. Therefore, it is crucial to identify and pay more attention to the most relevant modalities while ignoring the irrelevant ones. We have used and partially changed the attention network [48], which takes in both facial and speech modalities and outputs an attention score for each to ensure that only important modalities are emphasized. By assigning more weight to the modalities that contribute the most to the final prediction, this attention mechanism helps the model perform better by zeroing in on the most relevant information. Before being fed into the attention network, the feature vectors of both modalities are scaled to the same size. One approach to achieve this is by using a fully connected layer with a size of "s" to adjust the dimensionality of the feature vectors of all modalities. After normalization to size "s", the combined facial and speech feature set "A" may be written as $A = \left[A_f, A_s\right]$, where A_f denotes facial characteristics and A_s denotes speech features. Hence, A is a matrix with the dimensions "s by 2". We calculate the attention weight vector ω_f and the fused multimodal feature vector "F_{fs}" as follows:

$$X_{F_{fs}} = \tanh\left(W_{F_{fs}} A\right) \tag{4}$$

$$\omega_{fs} = softmax\left(w_{F_{fs}}^T X_{F_{fs}}\right) \tag{5}$$

$$F_{fs} = A\omega_{fs}^T \tag{6}$$

where $F_{fs} \in \mathbb{R}^s$, $\omega_{fs}^T \in \mathbb{R}^2$, $w_{F_{fs}} \in \mathbb{R}^s$, and $W_{F_{fs}} \in \mathbb{R}^{s\times s}$. After computing the fused multimodal feature vector F_{fs} using the attention weight vector ω_{fs}, we used it as an input to the classification layer to perform the final classification of multimodal emotions.

4. Experiment

4.1. Datasets

4.1.1. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [49]

The IEMOCAP dataset is a multimodal database containing audiovisual recordings of spontaneous dyadic interactions between human actors designed for research in multimodal emotion recognition. The dataset includes several modalities, such as audio recordings of speech and motion capture data and video recordings of facial expressions, head movements, and body gestures. The IEMOCAP dataset consists of five sessions conducted by ten distinct speakers, with each session containing recordings from two speakers, comprising one male and one female. The dataset is labeled with four main emotion classes (Figure 5): angry, sad, neutral, and happy. The total dataset consists of 4490 dialogues. There are angry (1103), sadness (1084), neutral (1708), and happy (595) emotional samples. To ensure a fair comparison during the final evaluation of our model using the IEMOCAP dataset, we implemented comprehensive end-to-end training by utilizing the original data reorganized by [20]. To be more specific (Figure 6), we allocated 70% and 10% of the data, which amounts to 3143 and 449 samples, as the training set and validation set, respectively. These samples were extracted from the first four sessions of the dataset, involving eight actors. The remaining 20% of the data, comprising 898 dialogues, was reserved as the test set. These dialogues specifically belong to Session 5 and involve two actors. In contrast to the approach taken in [50], we did not employ 10-fold cross-validation in our study. This decision was based on the impracticality of implementing cross-validation on deep learning models due to the substantial time and computational resources it would demand.

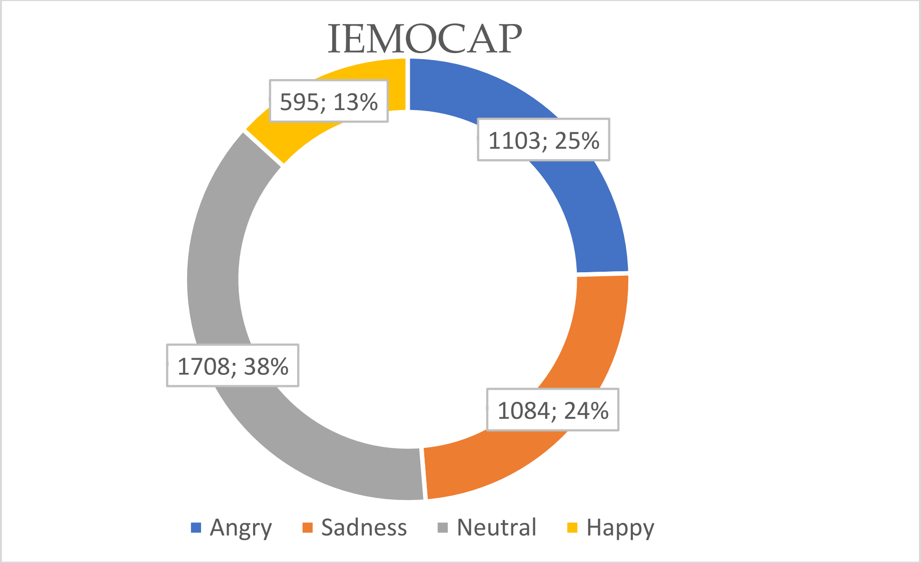


Figure 5. IEMOCAP data distribution.

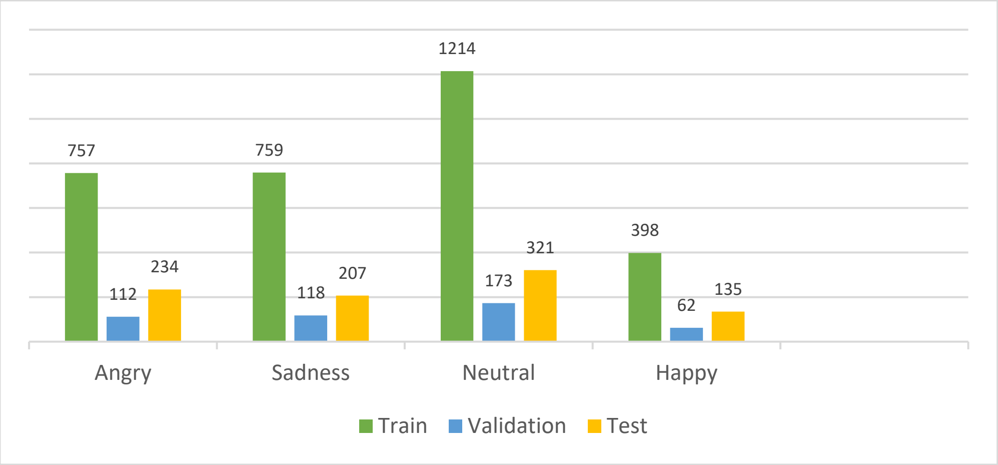


Figure 6. The train/validation and test data samples of the IEMOCAP dataset.

4.1.2. The CMU-MOSEI [51]

The CMU-MOSEI dataset is a large-scale, multimodal dataset for emotion analysis (Figure 7) in the context of conversational video data. The dataset contains more than 23,000 video clips from 1000 sessions with over 1200 participants. The video data is accompanied by speech transcripts, audio features, visual features, and labels indicating the valence and arousal levels. The CMU-MOSEI dataset comprises six emotion classes: anger, happiness, sadness, disgust, fear, and surprise. There are angry (4600), sadness (5601), disgust (3755), surprise (2055), happy (10752), and fear (1803) emotional samples. Similar to the division of the IEMOCAP dataset, we applied the same split to the CMU-MOSEI dataset (Figure 8). For the creation of training and validation sets, we allocated 70% and 10% of the data, amounting to 20,346 and 2411 samples, respectively. The remaining 20% of the data, comprising 5809 samples, was reserved specifically as the test set.

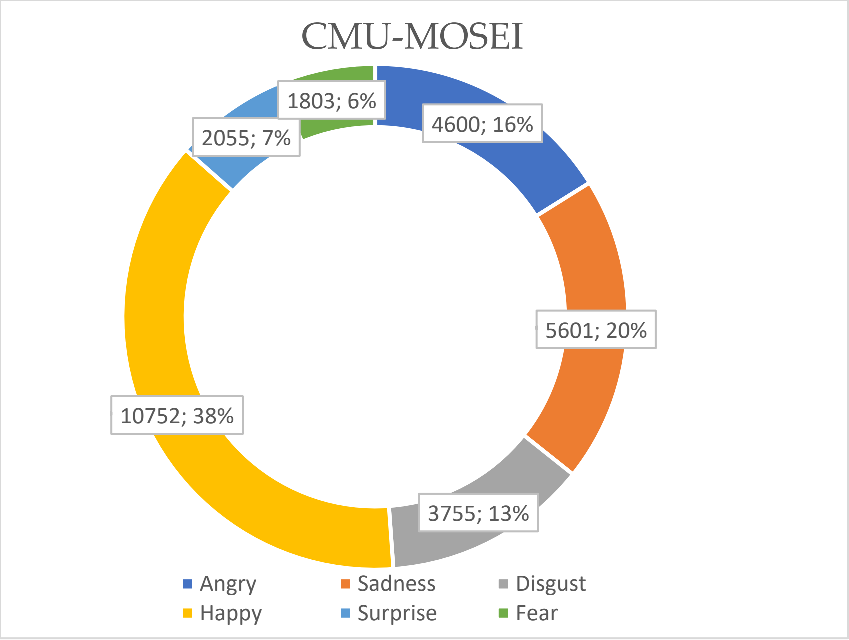


Figure 7. CMU-MOSEI data distribution.

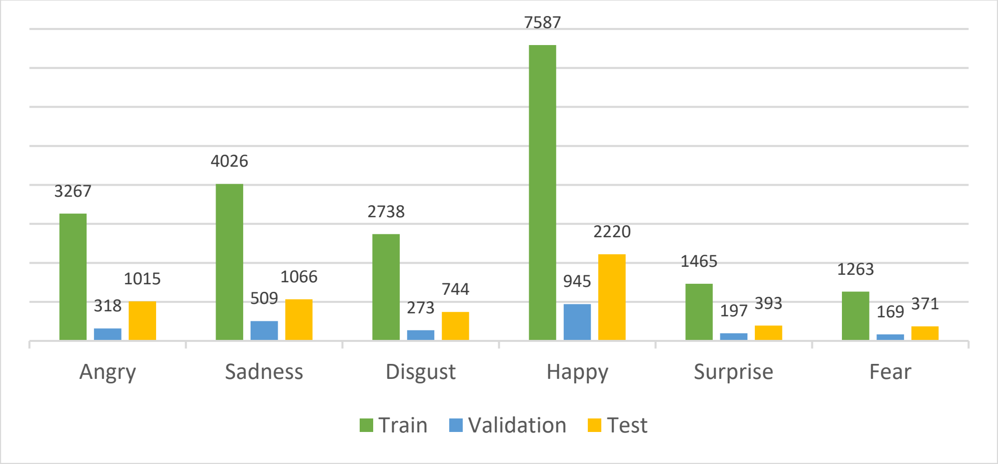


Figure 8. The train/validation and test data samples of the CMU-MOSEI dataset.

4.2. Evaluation Metrics

In our evaluation of the overall model performance, we employed several quantitative metrics to provide a comprehensive assessment. These metrics included the widely used F1 score and weighted accuracy (WA) to account for class imbalances and better capture the average performance across different classes.

WA is a valuable metric that takes into consideration the distribution of emotions within each class. It quantifies the ratio of correctly classified emotions to the total number of emotions belonging to a specific class. By considering the relative importance of each class, WA provides a more accurate representation of performance in scenarios where certain classes may have more instances than others.

By utilizing these metrics in our evaluation, we aimed to gain insights into the model's ability to accurately classify emotions across different classes, considering both individual class performance and overall class distributions. The formulas for the F1 score and WA metrics are as follows: [provide the formulas for the specific metrics].

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

$$\mathrm{WA} = \frac{\mathrm{TP} \times \mathrm{N}/\mathrm{P} + \mathrm{TN}}{2\mathrm{N}} \tag{8}$$

where N is the total number of negative labels and P is the total number of positive labels. TP represents true positives while TN represents true negatives.

4.3. Implementation Details

The implementation of the proposed approach involved the utilization of specific software, hardware configurations, and model parameters, as depicted in Table 1.

Table 1. Implementation settings.

Software	Programming tools	Python, Pandas, OpenCV, Librosa, Keras-TensorFlow
	OS	Windows 10
Hardware	CPU	AMD Ryzen Threadripper 1900X 8-Core Processor 3.80 GHz, TSMC, South Korea
	GPU	Titan Xp 32 GB
	RAM	128 GB

Table 1. *Cont.*

Parameters	Epochs	100
	Batch size	32
	Learning rate	0.001, Adam optimizer
	Regularization	L2 regularization, Batch normalization

4.4. Experimental Performance and Its Comparison

To provide an indication of the degree to which the proposed system is superior to those offered by the competition, we contrasted it with the criteria listed below. We selected the approaches for comparison based on the datasets. The results of our predictions are shown in Table 2. The selected and proposed systems are viable models for multimodal emotion recognition tasks. However, our system has outperformed the selected models in the MER tasks, particularly when combined with semantic information.

1. Wenliang et al. [20] completely constructed an end-to-end multimodal emotion recognition model that links and optimizes the two stages simultaneously.
2. Xi et al. [21]'s emotion identification from face video includes a semantic improvement module that guides the audio/visual encoder with text information, followed by a multimodal bottleneck transformer that reinforces audio and visual representations via cross-modal dynamic interactions.
3. Multimodal transformer [22]. The suggested technique intends to improve emotion identification accuracy by deploying a cross-modal translator capable of translating across three distinct modalities.

Table 2. Model comparisons on the CMU-MOSEI and IEMOCAP dataset.

Datasets	**Models**	**Metrics**	
		WA (%)	**F1 (%)**
IEMOCAP	Xia et al. [21]	–	64.6
	Multimodal transformer [22]	72.8	59.9
	Wenliang et al. [20]	-	57.4
	Our system	74.6	66.1
CMU-MOSEI	Xi et al. [21]	69.6	50.9
	Multimodal transformer [22]	80.4	67.4
	Wenliang et al. [20]	66.8	46.8
	Our system	80.7	73.7

Table 2 refers to the performance evaluation of a system on two different datasets—IEMOCAP and CMU-MOSEI. It is evident that the system surpasses those models in terms of performance. Specifically, on the IEMOCAP dataset, the system achieves a weighted accuracy WA of 74.6% and an F1 score of 66.1%. The WA is a metric that calculates the accuracy of predictions, taking into account the imbalance of the dataset, whereas F1 score is a harmonic mean of precision and recall, which measures the accuracy and completeness of the predictions. Similarly, on the CMU-MOSEI dataset, the system achieves a WA of 80.7% and F1 score of 73.7%, both of which are higher than the existing models evaluated on the same dataset.

According to Figure 9, the suggested system achieved 90.6% accuracy in WA on the CMU-MOSEI dataset for the happy emotion. It also had excellent accuracies for sadness and surprise, with 85.3% and 81.2%, accordingly. The model performed poorly in identifying angry and fear classes, with 73.9% and 71.0% accuracy, respectively. On the other hand, the proposed approach had the highest F1 rates for the emotions of happiness and disgust, earning 84.3% and 83.3%, respectively. Both surprise and fear were rated lower by the model on the F1 scale, with scores of 76.8% and 66.5%, for instance.

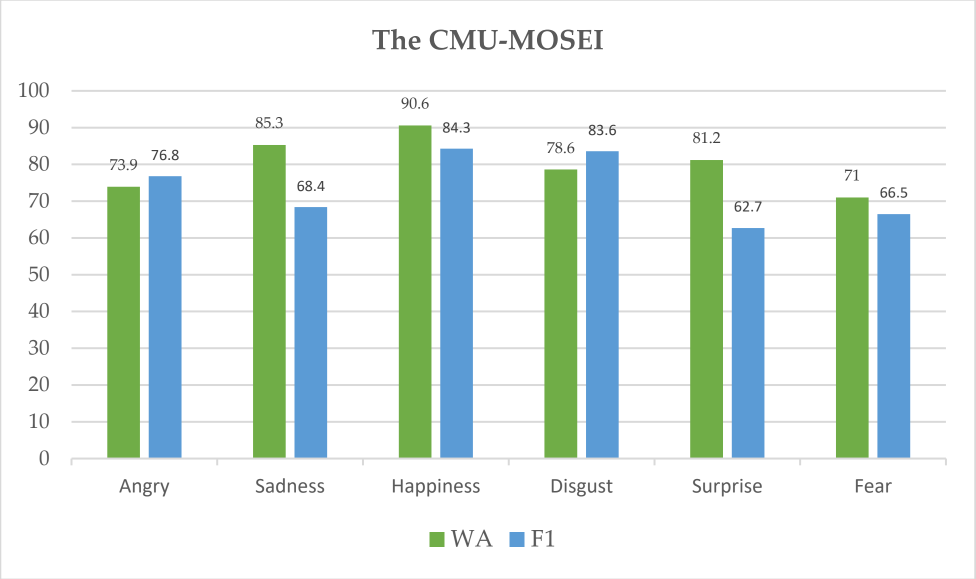


Figure 9. WA and F1 rates for each emotion class of the CMU-MOSEI dataset.

In the IEMOCAP dataset scenario, we can observe that the system performs inconsistently across distinct feelings. For instance, in Figure 10, when it comes to identifying the emotion "neutral", WA has a score of 83.1, suggesting that it does an excellent job. The system also shows effective recognition of sadness with a pretty high WA score of 77.9. Moreover, the system displays a relatively high level of accuracy in recognizing sadness (74.7) and angry (69.0) emotions, as shown by its high F1 scores. However, for the emotions of neutral and happiness, the system's F1 score is lower, with rates of 64.5 and 56.3, respectively. This suggests that the system may have more difficulty recognizing these emotions compared to the others.

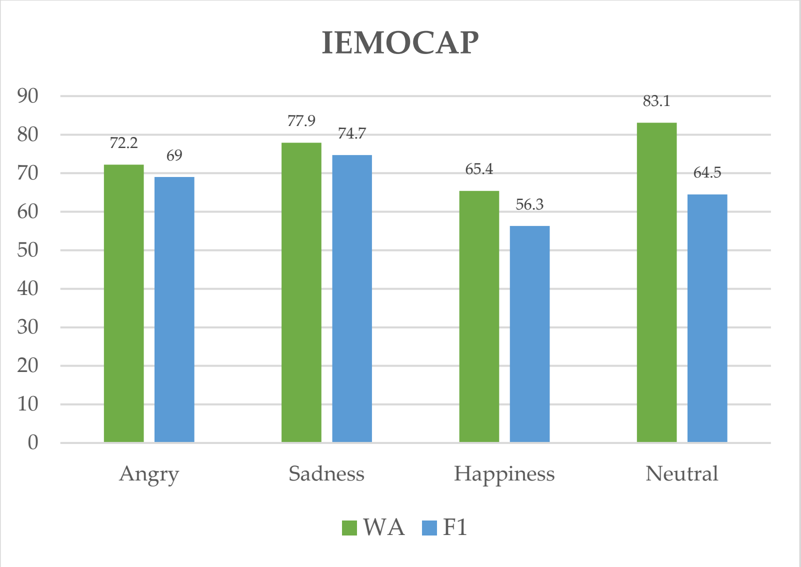
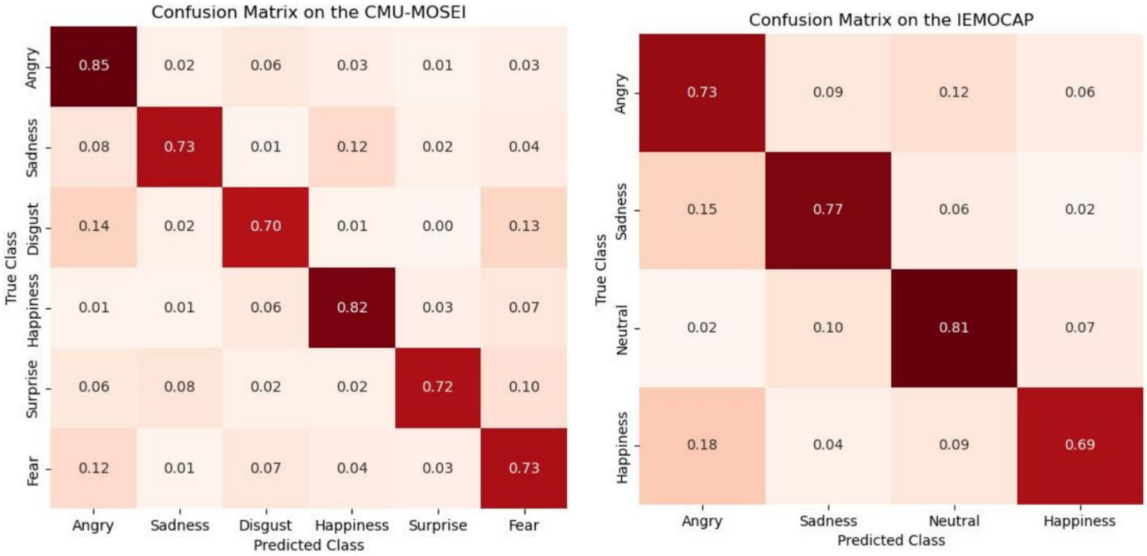


Figure 10. WA and F1 rates for each emotion class of the IEMOCAP dataset.

We evaluated the performance of the proposed model on the CMU-MOSEI and IEMOCAP datasets. The analysis was conducted using a confusion matrix, as illustrated in Figure 11. The model demonstrated an accuracy exceeding 71% for each emotion class, indicating a reasonably high level of classification accuracy. However, it is important to note that the evaluation dataset exhibited an imbalance in the distribution of samples among the emotion classes. This means that certain classes had a larger number of samples compared to others. Consequently, the model had a tendency to misclassify samples into the classes with a greater representation in the training data.



(a)

(b)

Figure 11. Confusion matrices on the (**a**) CMU-MOSEI and (**b**) IEMOCAP datasets.

Specifically, in the CMU-MOSEI dataset, the most common type of error occurred when samples were misclassified into the happy classes. This suggests that the model often labeled samples as happy, even if they belonged to a different emotion category. The prevalence of happy samples in the dataset likely influenced the model's predictions, resulting in this bias.

Similarly, in the IEMOCAP dataset, the most frequent misclassification involved samples being labeled as neutral. This indicates that the model tended to classify samples as neutral, regardless of their actual emotional label. The higher proportion of neutral samples in the dataset likely influenced the model's inclination to predict this class more often.

In summary, although the proposed algorithm achieved an overall satisfactory accuracy, the imbalanced distribution of samples across emotion classes led to a bias towards classes with a larger number of samples during the classification process. This highlights the importance of considering data balance and implementing strategies to address bias when training and evaluating emotion recognition models.

4.5. Discussion

In this research, we introduced an innovative approach for multimodal emotion recognition by integrating vocal and facial characteristics through an attention mechanism. Our proposed system overcomes the limitations of single-modality systems by combining facial expressions and speech features, leading to improved accuracy in recognizing emotions. By leveraging the valuable information from both modalities, our approach offers a more comprehensive and robust solution for emotion recognition.

The use of MFCCs in combination with a CNN enables our model to extract meaningful representations from the speech data, capturing both frequency-based and time-based attributes. This comprehensive representation facilitates the recognition of subtle variations and nuances in vocal expressions, leading to improved accuracy in identifying emotions. By effectively handling speech data of varying lengths, our system avoids the potential loss of important information that may occur when dealing with diverse utterances. This capability enhances the robustness and reliability of our emotion recognition system, ensuring that it can effectively capture and interpret the rich emotional cues present in speech signals.

Furthermore, our approach incorporates both low- and high-level facial features, extracted through a convolutional neural network (CNN), to achieve a comprehensive representation of facial expressions. The extraction of low-level facial features allows our model to capture intricate local details, such as subtle muscle movements, fine-grained changes in facial contours, and microexpressions. These minute details play a crucial role in conveying specific emotional cues. In addition to low-level features, our system also captures high-level facial features that encompass global facial expressions. These features encompass broader facial characteristics, including overall facial configurations, macroexpressions, and the interplay between different facial regions. By integrating high-level features, our system gains a holistic understanding of the facial expression as a whole, allowing it to capture the overall emotional state being conveyed.

The performance evaluation of the system is conducted on two specific datasets, namely IEMOCAP and CMU-MOSEI. While the results on these datasets show promising performance, it is essential to acknowledge that the system's effectiveness may vary when applied to other datasets. The generalizability and robustness of the system across a wider range of datasets should be further investigated.

The proposed system heavily relies on the attention mechanism to select the most important features from both facial and speech modalities. While this approach helps in focusing on informative parts, it introduces a potential vulnerability. The system's performance could be significantly affected if the attention mechanism fails to properly identify and assign appropriate weights to relevant features. Possible issues with attention mechanism performance and its impact on overall system accuracy should be considered. It is essential to consider and address any potential biases that might be present in the training data or the model itself. Biases in emotion recognition systems can arise due to imbalanced datasets, cultural or demographic biases, or biases inherent in the training process.

The superior performance of our system on these datasets highlights its potential in various applications such as affective computing, human–robot interaction, and mental health diagnosis. By leveraging both vocal and facial characteristics and employing an attention mechanism, our proposed methodology offers a promising approach for multimodal emotion recognition, contributing to advancements in the field.

5. Conclusions

In summary, emotions play a crucial role in human interactions, and there is a growing interest in multimodal emotion recognition that combines different modalities to provide a more comprehensive understanding of an individual's emotional state. However, recognizing emotions from a single modality is challenging, and deep neural networks have been used to extract the relevant features. Attention mechanisms have been shown to enhance the performance of deep neural networks by focusing on the informative parts of the input data.

This study proposes a novel multimodal emotion recognition system that integrates facial and speech features using an attention mechanism. By leveraging complementary information from both modalities, the proposed approach overcomes the limitations of unimodal systems and enhances emotion recognition accuracy. The proposed system for handling speech data of varying lengths utilizes time and spectral information in the speech modality, which enables models to concentrate on the most crucial parts of speech data and minimize the loss of important information. We have also utilized our previously proposed CNN model to acquire low- and high-level facial features. The generalizability of the system has been enhanced by mitigating the issue of overfitting. The effectiveness of the model is demonstrated on the IEMOCAP and CMU-MOSEI datasets; it has promising applications in areas such as affective computing, human–robot interaction, and mental health diagnosis.

Despite the success of this system, there are several challenges and opportunities for further research. For example, designing efficient and scalable attention mechanisms for large-scale datasets remains a major research direction, as well as integrating attention mechanisms with other techniques, such as reinforcement learning or meta-learning. Moreover, integrating personalized recommendation models [52,53] into multimodal emotion recognition will also be a future direction to significantly improve the emotional well-being and quality of life for individuals.

We believe that addressing these challenges and opportunities will facilitate more advanced and robust multimodal emotion recognition systems.

Author Contributions: D.M. and A.B.A. developed the method; D.M., A.B.A. and B.M. performed the experiments and analysis; D.M. wrote the paper; A.K. and T.K.W. supervised the study and contributed to the analysis and discussion of the algorithm and experimental results. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the GRRC program of Gyeonggi province. [GRRC-Gachon2021 (B03), Development of Healthcare Contents based on AI].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: D.M., A.B.A., A.K. and B.M. would like to express their sincere gratitude and appreciation to the supervisor, Taeg Keun Whangbo (Gachon University) for his support, comments, remarks, and engagement over the period in which this manuscript was written. Moreover, the authors would like to thank the editor and anonymous referees for the constructive comments in improving the contents and presentation of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Balci, S.; Demirci, G.M.; Demirhan, H.; Sarp, S. Sentiment Analysis Using State of the Art Machine Learning Techniques. In *Digital Interaction and Machine Intelligence, 9th Machine Intelligence and Digital Interaction Conference, Warsaw, Poland, 9–10 December 2021*; Lecture Notes in Networks and Systems; Biele, C., Kacprzyk, J., Kopeć, W., Owsiński, J.W., Romanowski, A., Sikorski, M., Eds.; Springer: Cham, Switzerland, 2022; Volume 440. [CrossRef]
2. Ahmed, N.; Al Aghbari, Z.; Girija, S. A systematic survey on multimodal emotion recognition using learning algorithms. *Intell. Syst. Appl.* **2023**, *17*, 200171. [CrossRef]
3. Gu, X.; Shen, Y.; Xu, J. Multimodal Emotion Recognition in Deep Learning:a Survey. In Proceedings of the 2021 International Conference on Culture-Oriented Science & Technology (ICCST), Beijing, China, 18–21 November 2021; pp. 77–82. [CrossRef]
4. Tang, G.; Xie, Y.; Li, K.; Liang, R.; Zhao, L. Multimodal emotion recognition from facial expression and speech based on feature fusion. *Multimedia Tools Appl.* **2022**, *82*, 16359–16373. [CrossRef]
5. Luna-Jiménez, C.; Griol, D.; Callejas, Z.; Kleinlein, R.; Montero, J.M.; Fernández-Martínez, F. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. *Sensors* **2021**, *21*, 7665. [CrossRef]
6. Sajjad, M.; Ullah, F.U.M.; Ullah, M.; Christodoulou, G.; Cheikh, F.A.; Hijji, M.; Muhammad, K.; Rodrigues, J.J. A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. *Alex. Eng. J.* **2023**, *68*, 817–840. [CrossRef]
7. Song, Z. Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory. *Front. Psychol.* **2021**, *12*, 759485. [CrossRef] [PubMed]
8. Abdusalomov, A.B.; Safarov, F.; Rakhimov, M.; Turaev, B.; Whangbo, T.K. Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm. *Sensors* **2022**, *22*, 8122. [CrossRef] [PubMed]
9. Hsu, J.H.; Su, M.H.; Wu, C.H.; Chen, Y.H. Speech emotion recognition considering nonverbal vocalization in affective conversations. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1675–1686. [CrossRef]
10. Ayvaz, U.; Gürüler, H.; Khan, F.; Ahmed, N.; Whangbo, T.; Bobomirzaevich, A.A. Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning. *Comput. Mater. Contin.* **2022**, *71*, 5511–5521. [CrossRef]
11. Makhmudov, F.; Mukhiddinov, M.; Abdusalomov, A.; Avazov, K.; Khamdamov, U.; Cho, Y.I. Improvement of the end-to-end scene text recognition method for "text-to-speech" conversion. *Int. J. Wavelets Multiresolution Inf. Process.* **2020**, *18*, 2050052. [CrossRef]

12. Vijayvergia, A.; Kumar, K. Selective shallow models strength integration for emotion detection using GloVe and LSTM. *Multimed. Tools Appl.* **2021**, *80*, 28349–28363. [CrossRef]
13. Farkhod, A.; Abdusalomov, A.; Makhmudov, F.; Cho, Y.I. LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. *Appl. Sci.* **2021**, *11*, 11091. [CrossRef]
14. Pan, J.; Fang, W.; Zhang, Z.; Chen, B.; Zhang, Z.; Wang, S. Multimodal Emotion Recognition based on Facial Expressions, Speech, and EEG. *IEEE Open J. Eng. Med. Biol.* **2023**, 1–8. [CrossRef]
15. Liu, X.; Xu, Z.; Huang, K. Multimodal Emotion Recognition Based on Cascaded Multichannel and Hierarchical Fusion. *Comput. Intell. Neurosci.* **2023**, *2023*, 1–18. [CrossRef] [PubMed]
16. Farkhod, A.; Abdusalomov, A.B.; Mukhiddinov, M.; Cho, Y.-I. Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces. *Sensors* **2022**, *22*, 8704. [CrossRef]
17. Chaudhari, A.; Bhatt, C.; Krishna, A.; Travieso-González, C.M. Facial Emotion Recognition with Inter-Modality-Attention-Transformer-Based Self-Supervised Learning. *Electronics* **2023**, *12*, 288. [CrossRef]
18. Krishna, N.D.; Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. *Interspeech* **2020**, *2020*, 4243–4247. [CrossRef]
19. Xu, Y.; Su, H.; Ma, G.; Liu, X. A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context. *Complex Intell. Syst.* **2023**, *9*, 951–963. [CrossRef]
20. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal end-to-end sparse model for emotion recognition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Online, 6–11 June 2021; pp. 5305–5316.
21. Xia, X.; Zhao, Y.; Jiang, D. Multimodal interaction enhanced representation learning for video emotion recognition. *Front. Neurosci.* **2022**, *16*, 1086380. [CrossRef]
22. Yoon, Y.C. Can We Exploit All Datasets? *Multimodal Emotion Recognition Using Cross-Modal Translation. IEEE Access* **2022**, *10*, 64516–64524. [CrossRef]
23. Yang, K.; Wang, C.; Gu, Y.; Sarsenbayeva, Z.; Tag, B.; Dingler, T.; Wadley, G.; Goncalves, J. Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1082–1097. [CrossRef]
24. Tashu, T.M.; Hajiyeva, S.; Horvath, T. Multimodal Emotion Recognition from Art Using Sequential Co-Attention. *J. Imaging* **2021**, *7*, 157. [CrossRef] [PubMed]
25. Kutlimuratov, A.; Abdusalomov, A.; Whangbo, T.K. Evolving Hierarchical and Tag Information via the Deeply Enhanced Weighted Non-Negative Matrix Factorization of Rating Predictions. *Symmetry* **2020**, *12*, 1930. [CrossRef]
26. Dang, X.; Chen, Z.; Hao, Z.; Ga, M.; Han, X.; Zhang, X.; Yang, J. Wireless Sensing Technology Combined with Facial Expression to Realize Multimodal Emotion Recognition. *Sensors* **2023**, *23*, 338. [CrossRef] [PubMed]
27. Nguyen, D.; Nguyen, D.T.; Sridharan, S.; Denman, S.; Nguyen, T.T.; Dean, D.; Fookes, C. Meta-transfer learning for emotion recognition. *Neural Comput. Appl.* **2023**, *35*, 10535–10549. [CrossRef]
28. Dresvyanskiy, D.; Ryumina, E.; Kaya, H.; Markitantov, M.; Karpov, A.; Minker, W. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. *Multimodal Technol. Interact.* **2022**, *6*, 11. [CrossRef]
29. Wei, W.; Jia, Q.; Feng, Y.; Chen, G.; Chu, M. Multi-modal facial expression feature based on deep-neural networks. *J. Multimodal User Interfaces* **2019**, *14*, 17–23. [CrossRef]
30. Gupta, S.; Kumar, P.; Tekchandani, R.K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia Tools Appl.* **2023**, *82*, 11365–11394. [CrossRef]
31. Chowdary, M.K.; Nguyen, T.N.; Hemanth, D.J. Deep learning-based facial emotion recognition for human–computer inter-action applications. *Neural Comput. Appl.* **2021**, 1–18. [CrossRef]
32. Li, J.; Zhang, X.; Huang, L.; Li, F.; Duan, S.; Sun, Y. Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network. *Appl. Sci.* **2022**, *12*, 9518. [CrossRef]
33. Kutlimuratov, A.; Abdusalomov, A.B.; Oteniyazov, R.; Mirzakhalilov, S.; Whangbo, T.K. Modeling and Applying Implicit Dormant Features for Recommendation via Clustering and Deep Factorization. *Sensors* **2022**, *22*, 8224. [CrossRef]
34. Zou, H.; Si, Y.; Chen, C.; Rajan, D.; Chng, E.S. Speech Emotion Recognition with Co-Attention Based Multi-Level Acoustic Information. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022. [CrossRef]
35. Kanwal, S.; Asghar, S.; Ali, H. Feature selection enhancement and feature space visualization for speech-based emotion recognition. *PeerJ Comput. Sci.* **2022**, *8*, e1091. [CrossRef] [PubMed]
36. Du, X.; Yang, J.; Xie, X. Multimodal emotion recognition based on feature fusion and residual connection. In Proceedings of the 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 24–26 February 2023; pp. 373–377. [CrossRef]
37. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN. *Int. J. Interact. Multimed. Artif. Intell.* **2021**, *6*, 112. [CrossRef]
38. Zhao, Y.; Guo, M.; Sun, X.; Chen, X.; Zhao, F. Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals. *IET Signal Process.* **2023**, *17*, e12201. [CrossRef]

39. Dobrišek, S.; Gajšek, R.; Mihelič, F.; Pavešić, N.; Štruc, V. Towards Efficient Multi-Modal Emotion Recognition. *Int. J. Adv. Robot. Syst.* **2013**, *10*, 53. [CrossRef]
40. Mamieva, D.; Abdusalomov, A.B.; Mukhiddinov, M.; Whangbo, T.K. Improved Face Detection Method via Learning Small Faces on Hard Images Based on a Deep Learning Approach. *Sensors* **2023**, *23*, 502. [CrossRef] [PubMed]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Makhmudov, F.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via At-tention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047. [CrossRef]
43. Araujo, A.; Norris, W.; Sim, J. Computing Receptive Fields of Convolutional Neural Networks. *Distill* **2019**, *4*, e21. [CrossRef]
44. Wang, C.; Sun, H.; Zhao, R.; Cao, X. Research on Bearing Fault Diagnosis Method Based on an Adaptive Anti-Noise Network under Long Time Series. *Sensors* **2020**, *20*, 7031. [CrossRef]
45. Hsu, S.-M.; Chen, S.-H.; Huang, T.-R. Personal Resilience Can Be Well Estimated from Heart Rate Variability and Paralinguistic Features during Human–Robot Conversations. *Sensors* **2021**, *21*, 5844. [CrossRef]
46. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
47. Ayetiran, E.F. Attention-based aspect sentiment classification using enhanced learning through cnn-Bilstm networks. *Knowl. Based Syst.* **2022**, *252*. [CrossRef]
48. Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; Morency, L.-P. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), Orleans, LA, USA, 18–21 November 2017; pp. 1033–1038. [CrossRef]
49. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation* **2008**, *42*, 335–359. [CrossRef]
50. Poria, S.; Majumder, N.; Hazarika, D.; Cambria, E.; Gelbukh, A.; Hussain, A. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. *IEEE Intell. Syst.* **2018**, *33*, 17–25. [CrossRef]
51. Zadeh, A.; Pu, P. Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), Melbourne, VIC, Australia, 15–20 July 2018; pp. 2236–2246.
52. Ilyosov, A.; Kutlimuratov, A.; Whangbo, T.-K. Deep-Sequence–Aware Candidate Generation for e-Learning System. *Processes* **2021**, *9*, 1454. [CrossRef]
53. Safarov, F.; Kutlimuratov, A.; Abdusalomov, A.B.; Nasimov, R.; Cho, Y.-I. Deep Learning Recommendations of E-Education Based on Clustering and Sequence. *Electronics* **2023**, *12*, 809. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.