*electronics*

*Article*

# Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders

**Fazliddin Makhmudov [1], Alpamis Kutlimuratov [2], Farkhod Akhmedov [1], Mohamed S. Abdallah [1,3] and Young-Im Cho [1,*]**

[1] Department of Computer Engineering, Gachon University, Seongnam 1342, Republic of Korea
[2] Department of AI and Software Engineering, Gachon University, Seongnam 13120, Republic of Korea
[3] Informatics Department, Electronics Research Institute (ERI), Cairo 11843, Egypt
* Correspondence: yicho@gachon.ac.kr

**Abstract:** Meticulous learning of human emotions through speech is an indispensable function of modern speech emotion recognition (SER) models. Consequently, deriving and interpreting various crucial speech features from raw speech data are complicated responsibilities in terms of modeling to improve performance. Therefore, in this study, we developed a novel SER model via attention-oriented parallel convolutional neural network (CNN) encoders that parallelly acquire important features that are used for emotion classification. Particularly, MFCC, paralinguistic, and speech spectrogram features were derived and encoded by designing different CNN architectures individually for the features, and the encoded features were fed to attention mechanisms for further representation, and then classified. Empirical veracity executed on EMO-DB and IEMOCAP open datasets, and the results showed that the proposed model is more efficient than the baseline models. Especially, weighted accuracy (WA) and unweighted accuracy (UA) of the proposed model were equal to 71.8% and 70.9% in EMO-DB dataset scenario, respectively. Moreover, WA and UA rates were 72.4% and 71.1% with the IEMOCAP dataset.

**Keywords:** speech emotion recognition; convolution neural network; attention; deep learning; modeling

**Citation:** Makhmudov, F.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047. https://doi.org/10.3390/electronics11234047

Academic Editor: Ruifeng Xu

Received: 26 October 2022
Accepted: 29 November 2022
Published: 6 December 2022

## 1. Introduction

Recently, the advancement of artificial intelligence (AI) has been triggering further development of the current human–machine communication trend. Particularly, speech emotion recognition (SER) has become increasingly important for researchers to understand and distinguish real human speech characters. Improvement in SER puts its applicable domains, which includes entertainment, monitoring drivers' behaviors, voice assistant, medicine, call centers, and online education, to the next stage of development. Current business domains utilize the effectiveness of SER through modern conversational systems, such as Google Assistant, Siri, and Alexa, to assist and attract their clients. Auspiciously recognizing emotional positions of people also helps to improve their business achievements. Feelings expressed via speech must be accurately identified and appropriately handled to provide more natural and transparent interactions between computers and people. However, building an effective SER is a complex and arduous effort owing to the utterance levels and abstract emotions [1]. Moreover, determining a methodologically felicitous algorithm is crucial in realizing and achieving a performance superior to that of the established benchmarks. Particularly, classic SER methods cover several concrete stages from inputting and preprocessing audio data to feature extraction and emotion classification (Figure 1).

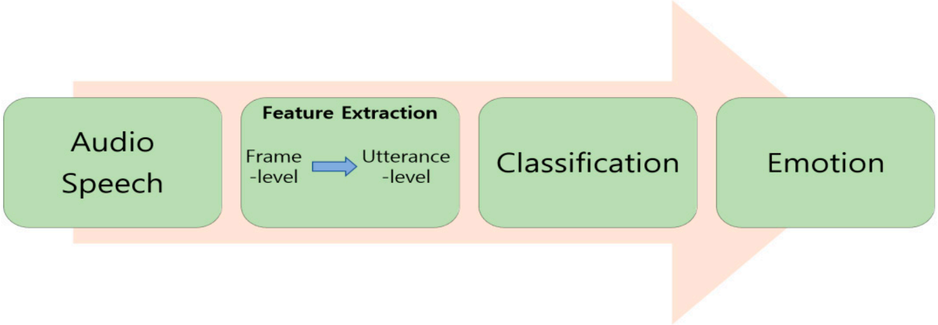


**Figure 1.** Steps of classic SER systems.

Support vector machine (SVM) [2,3], hidden Markov model (HMM) [4,5], Gaussian mixture model [6], and other methods [7,8] are representatives of classic SER methods. The classic methods typically begin by extracting numerous acoustic characteristics, including prosodic features (frame level, duration, energy, and pitch). Other important speech features, such as spectral, voice quality, and Teager energy operator (TEO) [9], are integrated to develop SER models using special mathematical functions. Nevertheless, the classic methods cause low accuracy performance, expensive computation, and effortful identification of various emotional cases [10]. Further, clarifying manually based on psychologists and sound specialists [11], neglecting time-domain features of raw audio data [12], and missing certain essential features during a feature extraction step [13] are common drawbacks of the classic SER models. However, the current trends of deep neural networks that are applied in the SER research field have demonstrated considerably remarkable efficiency and mitigated aforementioned problems. The automated extraction of emotional characteristics of the raw audio speech and understanding the relationships between those characteristics are the capabilities of learning algorithms in speech emotion identification. In comparison with classic approaches, it has demonstrated superior efficiency. For example, the authors of ref. [14] proposed a hybrid model based on long short-term memory (LSTM) and convolutional neural networks (CNNs) to apply time representation. However, despite their time representation, the duration of the speech was not managed and solved. Because, emotions may change with time during the continuous speech, and the changes may result in losing valuable information when an input speech data is segmented. On persistent SER, [15] developed a CNN model with two downsampling/upsampling structures and changing dilation factors of various layers. The changing factors may affect a model's performance and cause an overfitting problem. Therefore, to handle and lessen the problems, the model's complexity and number of parameters should be cut down. In addition, SER modeling with attention mechanism [11], transfer learning [16], and other deep neural networks [17,18] are illustrations of the current trends of AI. They mostly used high-level representations of speech features. Despite the advancement of the above mentioned DNN models, feature extraction and selection are the main parts of SER. Different speech features may offer a variety information on speech emotion. In addition to high-level representations, low-level (paralinguistic) feature representations may also have effects on detecting emotions and on how well a model works. Several studies use one and/or two sources of features to obtain optimized SER models. However, no experimentally approved appropriate set of features exist to build an effective SER model. Moreover, expensive computation and memory usage for deployment are the current limitations of the existing SER models that integrate multimodal features. An appropriate emotion detection model may be accomplished by designing effective research methodologies that integrate helpful speech emotions according to the methodology. Therefore, we attempted to build a novel SER model via attention-oriented parallel CNN encoders that parallelly acquires important features that are used for emotion classification by considering significant limitations of published methodologies and mitigates SER model performance in terms of accuracy. A minimized AlexNet CNN encoder for MFCC, fully convolutional network encoder for speech spectrogram, and CNN encoder for paralinguistic speech features were applied for their individual purposes and fed to attention mechanisms for further representation. The following are the contributions of this study:

- Improvement in terms of the model complexity
- Low-level (paralinguistic) feature representation
- Improvement in terms of model generalization
- Management of speeches of varying lengths
- A novel SER methodology that outperforms baseline models in terms of accuracy

The structure for the remainder of this article is as follows. The recent research works related to generating speech emotional features, pure deep learning SER models, and other DL methodologies that integrate attention mechanisms are specified in Section 2 of this article. In Sections 3 and 4, the proposed model is explained in detail, and the verification of its correctness is provided through empirical results and comparisons against benchmarks, respectively. Conclusions and future scope are presented in Section 5. Finally, the referenced literatures are cited, many of which are more recent articles.

## 2. Literature Overview

A study of emotions in speech is complicated and therefore requires a considerable effort by the researchers to algorithmically build a perfect model. Currently, various studies exist to detect people's emotions by analyzing speech features and effectively classify those features [19–21]. The steps involved in identifying emotions using raw speech data include selecting and extracting their features and then classifying the emotions based on the derived features. Therefore, an appropriate extraction of the features and their positive correlation have a significant impact on the performance level of the emotion classification model. In particular, modern SER models have benefited from the development of numerous innovative feature extraction approaches [22–26]. In ref. [22], modulation spectrum and frequency features were proposed with the help of amplitude and frequency modulation, and the two features were successfully integrated to the SER task. The authors of ref. [26] developed epoch-based features using a windowing technique and provided supplementary contribution to the SER model. The proposed research helped to increase the emotion recognition performance. Moreover, current trends and efficiency of deep learning models have advanced the SER task to the next level and conducted several studies regarding the task [18,27–30]. Research in ref. [27] suggested a new one-dimensional structure using LSTM and CNN algorithms to recognize emotions by showing an association of the semantic and spatial context of speech chunks. Furthermore, ref. [28] showed the role and strength of the speaker's temperament and mental health in SER by developing CNNs. Badshah et al. [30] used CNN based on rectangular kernels with different sizes to learn features, and the model outperformed baseline SER methods in terms of performance. In ref. [31], the authors proposed a model with a combination of SVM and deep belief network to classify emotions and to reveal speech features such as short-term energy, pitch, zero-crossing rate, formant, and MFCC. The authors in ref. [32] developed a hybrid DNN that comprises a binary weight network and recurrent neural network to detect common keywords in speech by considering to reduce energy consumption and space memory. In addition, several studies have been integrating attention mechanism with DNN for further development of latest deep learning models in SER [33–38]. In ref. [33], the authors aimed to solve pattern recognition, model parameters, and data sparsity issues; in addition, they derived spatial features by combining self-attention mechanism and a dilated CNN model. The authors in ref. [34] utilized the attention mechanism to improve prediction accuracy and the recursive feature elimination feature selection technique on prosodic speech features, including frequency, energy, and duration. Li et al. [37] suggested an attention pooling approach to recognize emotional representations where the representations are extracted from spectrograms of speech utterances using a deep CNN. The model in [38] fused the relationship between time and spatial features and intended to autonomously detect feature representations with the help of attention-based bidirectional LSTM recurrent neural networks and fully convolutional networks. Although authors in the aforementioned literature proposed their unique models for the SER task, certain limitations and low prediction accuracy challenges remain to be solved. In summary and to the best of our knowledge, no studies exist on parallel CNN encoders that integrate MFCC, paralinguistic, and speech spectrogram features to acquire a better model in terms of model complexity and accuracy. In addition, attention mechanisms that represent the encoded features for classification were not used. The following sections explain the entire workflow of the proposed model and prove the experimental results in detail.

## 3. Proposed Model

This section explains, in detail, the proposed model dedicated to the speech emotion recognition process. The model comprises exact components where each component has own effective contribution for emotion prediction. Figure 2 illustrates the entire modeling process. As illustrated, input speech data were divided into three speech features where subsequent model components parallelly encode those features. Then, the attention mechanism was used to receive and analyze the encoded features separately. Subsequently, all individual utterance features were concatenated as next flow to the consequent layer where the flow is trained for prediction. In the following subsections, each component of the proposed model is described in more depth.

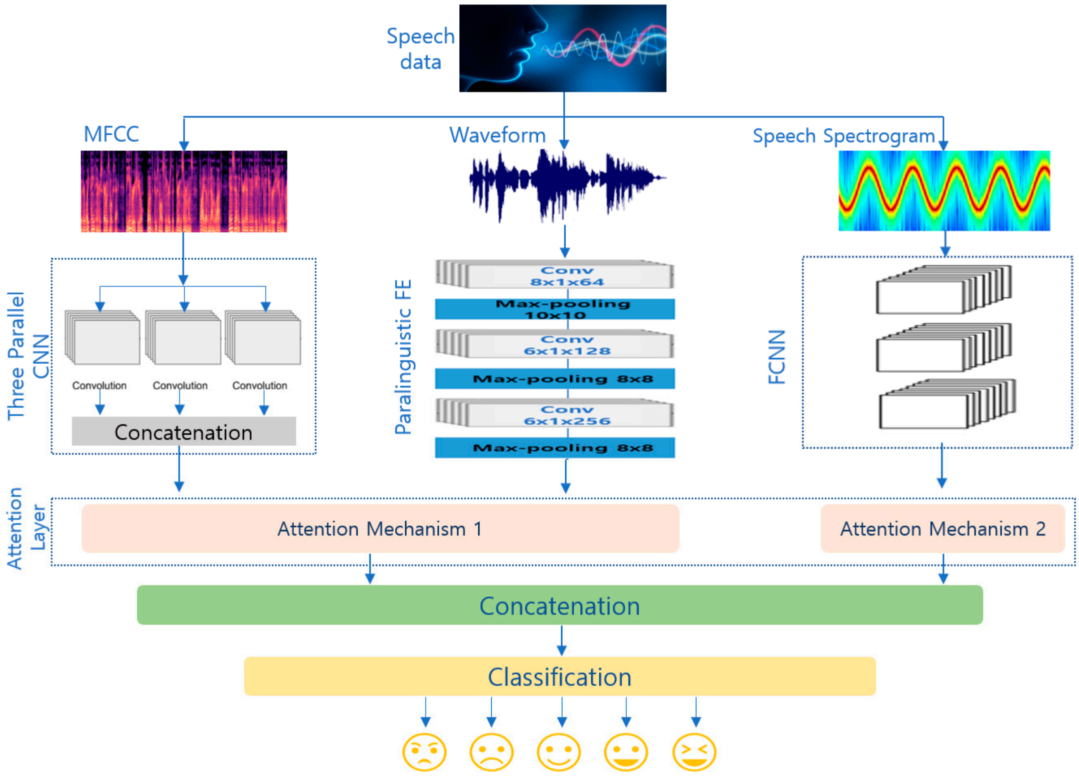


**Figure 2.** Workflow of the proposed model.

### *3.1. CNN-Based Feature Encoders*

#### 3.1.1. Parallel CNNs Encoder of MFCC

The parallel CNN encoders aim to stabilize the training phase of time and spectral data. Moreover, the larger CNN receptive field has a positive correlation with prediction accuracy [39]. However, a larger receptive field size increases the model parameters, resulting in model overfitting [40]. Considering the aforementioned objectives, we constructed the model components based on three CNNs that are located parallelly to derive different feature maps from MFCC via altered filter sizes, and the derived features were eventually concatenated and forwarded to the attention mechanism. In particular, the input speech data were normalized to compute the MFCC using a windowing technique to get 64-ms divided frames. Subsequently, every single 64 ms frame was subjected to Fourier transform. Then, CNN was trained by an initial 40 coefficients that were computed via inverse cosine transform on each MFCC frame. To construct the CNN-based three parallel encoders, the following approaches were employed:

- ➢ Deep CNN (adding more layers)
- ➢ Greater stride or average pooling
- ➢ Applying dilated convolutions
- ➢ Performing depth-wise convolutions

By appending extra layers to the CNN, convolution kernel size (Figure 3) expands the receptive field, thus leading to a deep CNN. Considering the model's overfitting problem on complex dimensions, the receptive field was computed [39] individually per dimension. Accordingly, to acquire time and spectral features, the convolution kernel size was 3 × 3 for the first CNN, 9 × 1 for the second CNN, and 1 × 11 for the third CNN. Because of this strategy, the computing complexity and number of model parameters of this component were decreased by $\frac{9\times11}{3\times3+9\times1+1\times11}$, compared with utilizing a single CNN-based encoder's identical receptive field size. Batch normalization (BN) receives feature-wise actions in the interpretation time where CNN's receptive field does not change. The receptive field of each particular layer is input speech data and its activations yield BN parameters. A convolutional kernel obtains "spots" from dilations. Because the amount of kernel weights remains identical, they do not exist anymore and are inapplicable in spatial neighboring of samples. A kernel diluted by a factor of "$\alpha$" generates "$\alpha$" striding in the samples that are applied for the convolution calculation process. Consequently, a kernel spatial length of "$k > 0$" will increase to "$\alpha(k-1)+1$", and layers that apply dilations utilize that extended spatial length [39]. Additionally, regarding channel or spatial dimensions, convolutions can also be distinguished. The distinguished convolutions and their analogous counterparts share the same receptive field characters. Particularly, the calculation process of the receptive field receives a kernel with size of "3" from distinguished 3 × 3 depth-wise convolution. Eventually, all derived encoded MFCC features from each CNN-based encoder were joined and sent to attention mechanism 1.

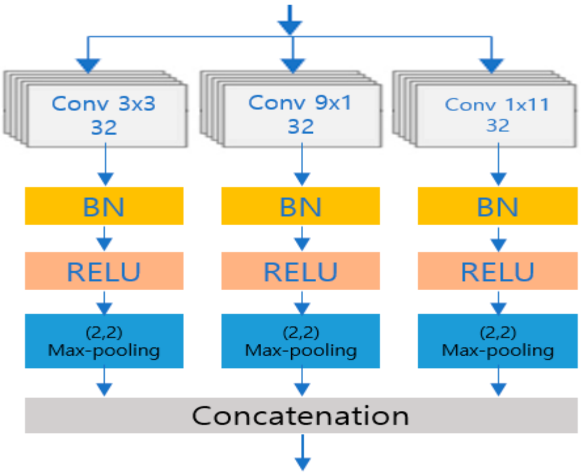


**Figure 3.** Parallel CNNs encoder of MFCC.

### 3.1.2. Paralinguistic Feature Encoder for Waveform (PFE)

To maintain the proposed model's training phase constancy and reinforce its generalization, the component attempts to occupy the proposed model with paralinguistic information. We believe that a better performance of the model may be achieved by combining several important features, as demonstrated by several developed models [41,42]. The PFE encoder includes three consecutive convolution layers. The computation of the convolutional layer is performed based on Equation (1) where $f(x)$ represents the kernel function that is conducted on input waveform data.

$$(f * h)(t) = \sum_{k=-T}^{T} f(t)h(t-k) \tag{1}$$

The input waveform data were first preprocessed to obtain unit variance and zero mean, and then, they were divided into 20 s intervals and employed as inputs to the convolutional layer. Max-pooling operation was then performed to decrease the dimensionality. The convolution kernel size impacts the selection of an effective pooling size, and an empirically basic approach was selected as expressed in Equation (2).

$$OR = \frac{KS - 1}{KS + PS - 1} \tag{2}$$

Here, $PS$ denotes the pooling size, $KS$ denotes the kernel size, and $OR$ denotes the overlap rate. Evidently, the $OR$ should be less than 1, and generally, it is believed to be around 0.5 when making hand-crafted features. Strides consider the complete information, whereas max-pooling just considers the most crucial information and discards the irrelevant data. Therefore, the $OR$ must be maintained below 0.5 to prevent it from obtaining the same characteristics for subsequent frames. To construct the PFE structure, as illustrated in Figure 4, the factor for $OR$ is considered, and the value is equal to "0.35".

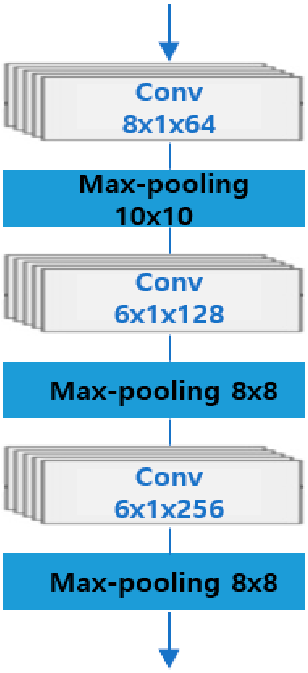


**Figure 4.** Structure of a PFE encoder.

Computations on max-pooling and convolution were performed with greater and tiny strides and kernel sizes, respectively. *Convolution layer 1*: To obtain features from the segmented input waveform data, 64 temporal filters were applied with a $KS$ of "8". *Max-pooling layer 1*: To reduce the signal's frame rate and maintain the most meaningful characteristics, max-pooling was utilized with a size of 10, which is identical to the $KS$ of the first convolutional layer. *Convolution layer 2*: Channel and kernel sizes in this layer were "128" and "6", respectively, aiming to derive more high-level representations. *Max-pooling layer 2*: With a size of "8", the layer is pooled by considering the $OR$ factor that should be below "0.5". *Convolution layer 3*: More high-level representations than the previous convolution layer are obtained from last convolution layer 3 where filter and kernel sizes were "256" and "6", respectively. *Max-pooling layer 3*: In the end, max-pooling operation was performed over time domain with a size of "8". Eventually, the encoded waveform feature output from the PFE encoder component of the proposed model was sent to attention mechanism 1.

### 3.1.3. Fully Convolutional Network Encoder of Spectrogram

The aim of this component of the proposed model was to prevent the loss of important information, and here, a fully convolutional neural network (FCNN) was used to reach that goal. The FCNN does not require any process for segmentation to deal with different lengths of speech data. Moreover, several deep-learning based models [43–46] were developed to build effective utterance features and achieve better accuracy results. In ref. [45], specific-sized chunks were derived by dividing raw speech spectrograms to meet CNN specifications. Therefore, every segmented utterance chunk received the emotion description of the relevant entire utterance. However, the process is not entirely logical to believe that the entire emotion does not express its meaning in segmented utterance chunks. We believe that the entire process of dividing speech spectrogram into chunks leads to loss of speech coherence that represents an altering emotion. Therefore, FCNN is integrated as a component of the proposed model to mitigate losing information and to manage different lengths of speech spectrogram. Furthermore, FCNN can process input data of any size and create appropriate interpreted and learned output. In particular, the FCNN was extracted from the AlexNet [44], which lost all fully linked layers, and it served as the component's encoder (Figure 5) of the proposed model.

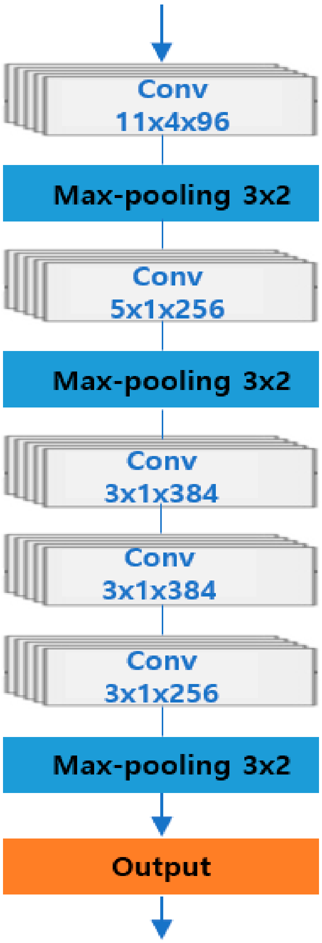


**Figure 5.** Structure of an FCNN encoder.

The FCNN comprises five convolutional layers in which a "local response normalization" was used after the first and second layers, and an activation function "ReLU" was used after each convolutional layer. Avoiding the saturation of ReLU results from its advantageous quality, i.e., it does not really require the normalization of input data. In case the ReLU receives positive feeds from certain training datasets, learning will occur within that neuron. However, we attempted to deliver generalization using a local normalization approach as expressed in Equation (3).

$$b^{i}_{x,y} = \frac{a^{i}_{x,y}}{\left(k + \alpha \sum_{j=\max(0,\, i-\frac{n}{2})}^{\min(N-1,\, i+\frac{n}{2})} \left(a^{j}_{x,y}\right)^{2}\right)^{\beta}} \tag{3}$$

where the constant values of $\beta = 0.75,\ n = 5,\ k = 2,$ and $\alpha = 0.0001$ are used as hyperparameters of the local response normalization N—all kernels of the corresponding layer. Settings of convolutional layers are indicated as follows: ("kernel size" × "*stride size*" × "*channels*"). The settings of the first convolutional layer are 11 × 4 × 96, and the parameters of the second and third convolutional layers are 5 × 1 × 256 and 3 × 1 × 384, respectively, which receive input after local response normalization and pooling. The fourth layer receives parameters identical to those of the third layer. The parameters of the fifth layer are 3 × 1 × 256. The results of adjacent neuronal groups in the same kernel map are summarized by pooling layers of CNNs. The FCNN encoder provides three-dimensional array of X × Y × Z with their respective sizes. The spectrogram's time and frequency domains are expressed by the characters "Y" and "X", respectively, and the size of the channel is represented by "Z". A set with a different length that includes "k" components is supposed as the output K = Y × X. Each component "K" is a Z-dimensional vector that encodes a particular segment of input speech spectrogram and is expressed by Equation (4).

$$\mathrm{C} = \{\mathrm{c}_1, \cdots \mathrm{c}_k\},\ \mathrm{c}_i \in \mathbb{R}^{\mathrm{Z}} \tag{4}$$

Eventually, encoded speech feature output from the FCNN encoder component of the proposed model is sent to attention mechanism 2.

### *3.2. Attention Mechanisms*

#### 3.2.1. Attention Mechanism "1"

Attention mechanism "1" was performed with the help of concatenation and attention mechanism techniques. The concatenation technique is a simple feature merging function where encoded MFCC feature $\mathrm{ef}_{mfcc} \in \mathbb{R}^{d_{mfcc}}$ and waveform feature $\mathrm{ef}_{wave} \in \mathbb{R}^{d_{wave}}$ are concatenated. This can be mathematically expressed as follows: $e_{con} = \left[\mathrm{ef}_{mfcc},\ \mathrm{ef}_{wave}\right]$.

To accomplish the attention mechanism, linear projection was applied on both the features to place them in the same dimension "$d_{mw}$".

$$\begin{gathered} \widetilde{\mathrm{ef}}_{mfcc} = \mathrm{P}_{mfcc}\mathrm{ef}_{mfcc} + \mathrm{c}_{mfcc} \\ \widetilde{\mathrm{ef}}_{wave} = \mathrm{P}_{wave}\mathrm{ef}_{wave} + \mathrm{c}_{wave} \end{gathered} \tag{5}$$

$\mathrm{P}_{mfcc} \in \mathbb{R}^{d_{fw} \times d_{mfcc}}$ and $\mathrm{P}_{wave} \in \mathbb{R}^{d_{fw} \times d_{wave}}$ indicate projection matrices of both features.

$$\begin{gathered} att1 = \alpha_{mfcc}\widetilde{\mathrm{ef}}_{mfcc} + \alpha_{wave}\widetilde{\mathrm{ef}}_{wave} \\ \alpha_j = softmax\left(\frac{\widetilde{ef}_j z_j}{\sqrt{d_{mw}}}\right) \end{gathered} \tag{6}$$

where $z \in \mathbb{R}^{d_{mw}}$ is the learnable vector of both features.

### 3.2.2. Attention Mechanism "2"

Each component of C and time-frequency entities may not evenly contribute to the emotional classes. Therefore, an attention mechanism technique was applied to capture crucial components to the emotion utterance, and then they were combined to create its vector. The attention technique was accomplished through Equations (7)–(9).

$$w_i = v^T \tan\text{h}(Wc_i + b) \tag{7}$$

$$\alpha_i = \frac{\exp(\beta w_i)}{\sum_{k=1}^{L} \exp(\beta w_k)} \tag{8}$$

$$ue = \sum_{i=1}^{L} \alpha_i c_i \tag{9}$$

To construct a distinct form of $c_i$, it was sent to multilayer perceptron based on the tanh activation function owing to its nonlinearity. Subsequently, the relevance weight ($w_i$) was acquired by calculating the inner product between learnable vector *"v"* and constructed distinct vector. Next, normalized relevance weight "$\alpha_i$" was computed using the softmax function. Eventually, emotion utterance vector "*ue*" was acquired with the calculation of sum of C.

## 4. Empirical Veracity and Discussion

### *4.1. Datasets*

IEMOCAP and EMO-DB datasets were used to evaluate the proposed model and prove its efficacy over counterparts. The datasets are publicly available and important ones for research community to model and study in speech emotional recognition task.

### 4.1.1. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database

The IEMOCAP dataset [47] offers observations of real emotive interactions and comprises dialogue (audio) and sentence (text) data. The actors and actresses offered a variety of improvised imaginary situations that were intended to evoke various emotions. We used only speech utterances (3784 samples) in dialogue format that depict four main emotion classes of angry, sad, neutral, and happy. The emotion distribution of the IEMOCAP dataset is shown in Figure 6.

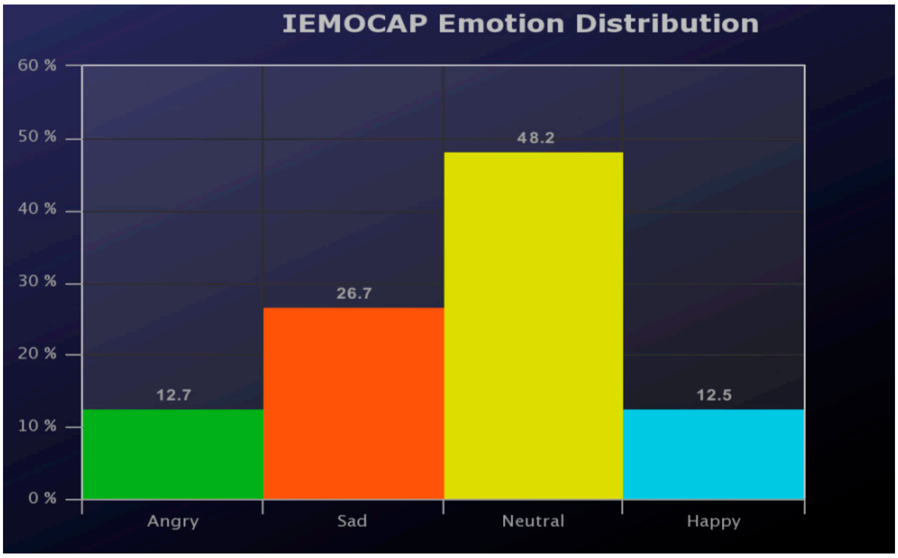


**Figure 6.** IEMOCAP emotion data distribution.

#### 4.1.2. EMO-DB: Berlin Emotional Database

The German emotional data, or EMO-DB [48], are publicly accessible. Seven emotion classes (disgust, sadness, fear, boredom, anger, neutral, and joy) are represented by the 535 overall utterances (samples) in the dataset. A wav file commonly lasts for 3 s. The emotion distribution of the EMO-DB dataset is shown in Figure 7.

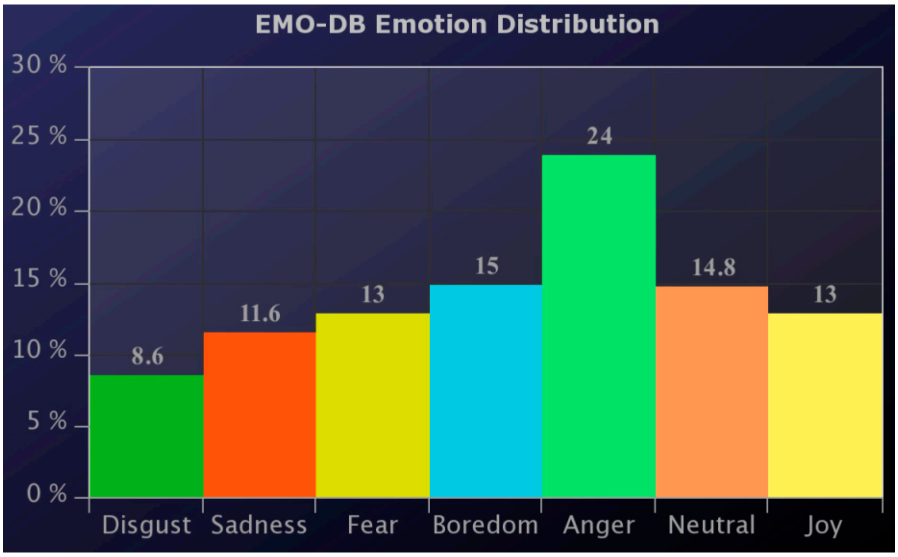


**Figure 7.** EMO-DB emotion data distribution.

The emotional classes in both datasets were unbalanced, and all empirical veracities were taken with five-fold cross validation. Each dataset had a training set randomly made of 80% of the data and a test set made of 20% of the data. A 16 kHz sampling rate was used for the input speech data. Because the datasets were unbalanced, weighted accuracy (WA) and unweighted accuracy (UA) were used as evaluation metrics. The ultimate outcome was determined by averaging all the empirical findings.

### *4.2. Software and Hardware Configuration*

The proposed approach was implemented using the following software and hardware configurations, as presented in Table 1.

**Table 1.** Software and hardware configuration.

| | | |
|---|---|---|
| Software | Programming tools | Python, Pandas,<br>Keras-TensorFlow |
| | OS | Windows 10 |
| Hardware | CPU | AMD Ryzen Threadripper 1900X 8-Core Processor 3.80 GHz |
| | GPU | Titan Xp 32 GB |
| | RAM | 128 |

The experiments were conducted on Titan Xp 32 GB for 300 epochs and 32 batch sizes, 128 GB RAM, AMD Ryzen Threadripper 1900X 8-Core with OS Windows 10. The proposed model was trained and tested via hyperparameters and the Adam optimizer with a learning rate of $10^{-4}$. The learning rate was divided by 10 every 20 epochs.

### *4.3. Model Performance and Its Comparisons*

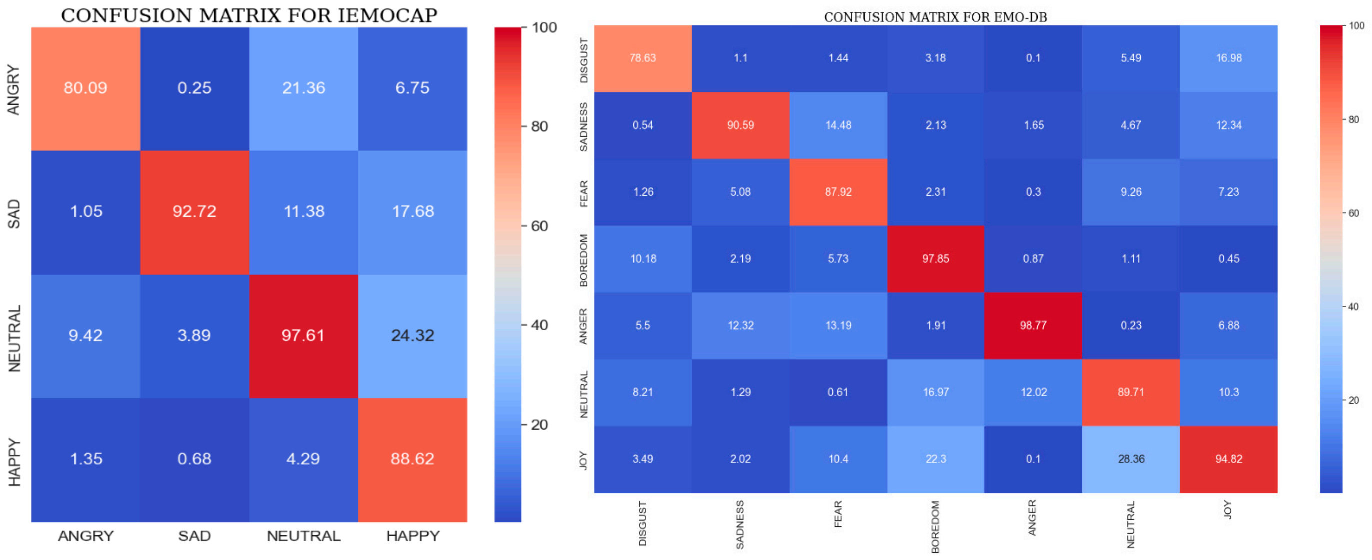
To indicate the extent to which the suggested approach is better than the competitors, we compared it with the following benchmarks, and the prediction results are presented in Table 2.

1. BLSTM-FCN two-layer attention [38]: Attention-aware BLSTM-RNN and FCN networks to learn spatial and temporal representations to predict emotions.
2. Deep CNN 2D [30]: A deep CNN model to derive discriminative features that uses rectangular kernels of varying shapes and sizes, along with max pooling in rectangular neighborhoods.
3. ATFNN [49]: Attentive Time-Frequency Neural Network (ATFNN) to learn the discriminative speech emotion feature for SER.
4. SFE [50]: the model aims to design and implement a novel feature extraction method that can extract features to recognize different emotions.

**Table 2.** Prediction performance and comparisons.

| Models | EMO-DB | | IEMOCAP | |
|---|---|---|---|---|
| | WA in % | UA in % | WA in % | UA in % |
| BLSTM-FCN two-layer attention | - | - | 68.1 | 67 |
| Deep CNN 2D | 69.2 | 67.9 | 71.2 | 70.6 |
| SFE | 71.1 | 68.4 | - | - |
| ATFNN | - | - | 72.66 | 64.48 |
| **Proposed model** | **71.8** | **70.9** | **72.4** | **71.1** |

The proposed model was compared with the selected benchmark deep learning-based models. In both datasets, the proposed model achieved better weighted accuracy (WA) and unweighted accuracy (UA) as presented in Table 3 among the models and confusion matrices presented in Figure 8. Especially, the WA and UA of the proposed model were equal to 71.8% and 70.9% in the EMO-DB dataset scenario, respectively. Moreover, the WA and UA rates were 72.4% and 71.1% with the IEMOCAP dataset.



(**a**) (**b**)

**Figure 8.** Confusion matrices on (**a**) the IEMOCAP and (**b**) EMO-DB datasets.

**Table 3.** Average accuracy comparisons between models.

| Models | Average Accuracy | |
|---|---|---|
| | **EMO-DB** | **IEMOCAP** |
| BLSTM-FCN<br>two-layer attention | 86.54 | - |
| Deep CNN 2D | 82.20 | 78.61 |
| SFE | 80.32 | - |
| ATFNN | 87.5 | 89.1 |
| **Proposed model** | **89.76** | **91.18** |

Table 3 presents a comparison between the emotion recognition accuracy of the proposed model and selected benchmarks.

## 5. Conclusions and Future Scope

The development of a methodologically felicitous algorithm to acquire valuable speech features is a crucial task to achieve better model performance over established benchmarks. Obtaining and interpreting speech features to recognize emotions in speech data might also be difficult. The main factors that make it hard to design a SER model are extracting useful features and classifying them correctly. However, the advancement of modern deep learning algorithms has been mitigating these challenging actions. Therefore, we used attention-oriented parallel CNN encoders that obtain important features at the same time and those used to classify emotions to come up with a new SER model. The core of the proposed model relies on CNN encoders for speech spectrogram, paralingusitic characteristics and MFCC, and attention mechanisms for further representation and classification. Our model allows managing different-length speeches, representing low-level (paralinguistic) features. The attention mechanisms make it possible for the network to concentrate on the emotional parts of the acquired features. Empirical veracity was executed on EMO-DB and IEMOCAP open datasets, and the results showed that the proposed model was more efficient than the baseline models. Despite producing improved recognition outcomes, the proposed model has certain limitations that may be solved in the future work. Namely, the better performances of the proposed model were obtained only with EMO-DB and IEMOCAP datasets. Therefore, we would like to obtain and integrate different visual and audio features, and modify the architecture of the proposed model to apply for other existing datasets. Furthermore, numerous potential future research areas have been identified from the current study and may involve the following:

- integrating EGG signals through deep learning algorithms
- combining emotions into a recommendation system [51,52]
- analyzing emotions of visually impaired people [53]
- To enhance speech emotion characteristics' resilience and discriminability [49]

These types of future studies might be attractive for employment in various other speech datasets.

**Author Contributions:** This manuscript was designed and written by F.M., A.K., F.A., M.S.A. and Y.-I.C. 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 research was supported by the MSIT (Ministry of Science and ICT), Republic of Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2017-0-01630) supervised by the IITP (Institute for Information & communications Technology Promotion) and by the Gachon University research fund of 2020 (GCU-202004350001).

**Acknowledgments:** The authors F.M., A.K., F.A. and M.A would like to express their sincere gratitude and appreciation to the supervisor, Young Im Cho (Gachon University) for her 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. Zhang, Y.; Du, J.; Wang, Z.; Zhang, J.; Tu, Y. Attention Based Fully Convolutional Network for Speech Emotion Recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 1771–1775. [CrossRef]
2. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [CrossRef]
3. Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P.; Tan, G.-Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing* **2018**, *273*, 271–280. [CrossRef]
4. Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov Model based speech emotion recognition. In Proceedings of the International Conference on Multimedia & Expo, Baltimore, MD, USA, 6–9 July 2003. [CrossRef]
5. New, T.L.; Foo, S.W.; Silva, L.C.D. Classification of stress in speech using linear and nonlinear features. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 Proceedings (ICASSP '03), Hong Kong, China, 6–10 April 2003; Volume 2, p. II-9.
6. Koolagudi, S.G.; Murthy, Y.V.S.; Bhaskar, S.P. Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. *Int. J. Speech Technol.* **2018**, *21*, 167–183. [CrossRef]
7. Henríquez, P.; Alonso, J.B.; Ferrer, M.A.; Travieso, C.M.; Orozco-Arroyave, J.R. Nonlinear dynamics characterization of emotional speech. *Neurocomputing* **2014**, *132*, 126–135. [CrossRef]
8. Milton, A.; Roy, S.S.; Selvi, S.T. SVM scheme for speech emotion recognition using mfcc feature. *Int. J. Comput. Appl.* **2013**, *69*, 34–39. [CrossRef]
9. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
10. An, X.D.; Ruan, Z. Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. *J. Phys. Conf. Ser.* **2021**, *1861*, 012064. [CrossRef]
11. Zhang, Z.; Wu, B.; Schuller, B. Attention-augmented End-to-end Multi-task Learning for Emotion Prediction from Speech. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6705–6709. [CrossRef]
12. 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 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022.
13. Zhang, H.; Gou, R.; Shang, J.; Shen, F.; Wu, Y.; Dai, G. Pre-trained Deep Convolution Neural Network Model with Attention for Speech Emotion Recognition. *Front. Physiol.* **2021**, *12*, 643202. [CrossRef] [PubMed]
14. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 5200–5204.
15. Khorram, S.; Aldeneh, Z.; Dimitriadis, D.; McInnis, M.; Provost, E.M. Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition. *arXiv* **2017**, arXiv:1708.07050.
16. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 478–484.
17. Lech, M.; Stolar, M.; Best, C.; Bolia, R. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Front. Comput. Sci.* **2020**, *2*, 14. [CrossRef]
18. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323. [CrossRef]
19. 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]
20. Tripathi, S.; Kumar, A.; Ramesh, A.; Singh, C.; Yenigalla, P. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. *arXiv* **2019**, arXiv:1906.05681. [CrossRef]
21. Atmaja, B.T.; Sasou, A. Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations. *Sensors* **2022**, *22*, 6369. [CrossRef] [PubMed]
22. Kerkeni, L.; Serrestou, Y.; Raoof, K.; Mbarki, M.; Mahjoub, M.A.; Cleder, C. Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Commun.* **2019**, *114*, 22–35. [CrossRef]
23. Gong, Y.; Chung, Y.-A.; Glass, J. AST: Audio Spectrogram Transformer. *arXiv* **2021**, arXiv:2104.01778.
24. Guo, S.; Feng, L.; Feng, Z.B.; Li, Y.H.; Wang, Y.; Liu, S.L.; Qiao, H. Multi-view laplacian least squares for human emotion recognition. *Neurocomputing* **2019**, *370*, 78–87. [CrossRef]
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. Fahad, M.S.; Deepak, A.; Pradhan, G.; Yadav, J. DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features. *Circuits Syst. Signal Process.* **2021**, *40*, 466–489. [CrossRef]
27. Mustaqeem; Kwon, S. CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network. *Mathematics* **2020**, *8*, 2133. [CrossRef]
28. Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G. Continuous Speech Emotion Recognition with Convolutional Neural Networks. *J. Audio Eng. Soc.* **2020**, *68*, 14–24. [CrossRef]
29. Shrestha, L.; Dubey, S.; Olimov, F.; Rafique, M.A.; Jeon, M. 3D Convolutional with Attention for Action Recognition. *arXiv* **2022**, arXiv:2206.02203. [CrossRef]
30. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **2019**, *78*, 5571–5589. [CrossRef]
31. Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. *Sensors* **2017**, *17*, 1694. [CrossRef]
32. Liu, B.; Qin, H.; Gong, Y.; Ge, W.; Xia, M.; Shi, L. EERA-ASR: An Energy-Efficient Reconfigurable Architecture for Automatic Speech Recognition with Hybrid DNN and Approximate Computing. *IEEE Access* **2018**, *6*, 52227–52237. [CrossRef]
33. Mustaqeem; Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.* **2021**, *102*, 107101. [CrossRef]
34. Alex, S.B.; Mary, L.; Babu, B.P. Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. *Circuits Syst. Signal Process.* **2020**, *39*, 5681–5709. [CrossRef]
35. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems, Proceedings of the Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015*; MIT Press: Cambridge, MA, USA, 2015; pp. 577–585.
36. Abdusalomov, A.; Baratov, N.; Kutlimuratov, A.; Whangbo, T.K. An Improvement of the Fire Detection and Classification Method Using YOLOv3 for Surveillance Systems. *Sensors* **2021**, *21*, 6519. [CrossRef]
37. Li, P.; Song, Y.; McLoughlin, I.; Guo, W.; Dai, L. An attention pooling based representation learning method for speech emotion recognition. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3087–3091. [CrossRef]
38. Zhao, Z.; Bao, Z.; Zhao, Y.; Zhang, Z.; Cummins, N.; Ren, Z.; Schuller, B. Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition. *IEEE Access* **2019**, *7*, 97515–97525. [CrossRef]
39. Araújo, A.F.; Norris, W.; Sim, J. Computing Receptive Fields of Convolutional Neural Networks. *Distill* **2019**, *4*, e21. [CrossRef]
40. 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] [PubMed]
41. 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]
42. 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; IEEE: Piscataway, NJ, USA, 2017; pp. 2227–2231.
43. Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.A.; Alhadlaq, A.; Lee, H.-N. Two-Way Feature Extraction for Speech Em otion Recognition Using Deep Learning. *Sensors* **2022**, *22*, 2378. [CrossRef]
44. Mocanu, B.; Tapu, R.; Zaharia, T. Utterance Level Feature Aggregation with Deep Metric Learning for Speech Emotion Recognition. *Sensors* **2021**, *21*, 4233. [CrossRef] [PubMed]
45. Satt, A.; Rozenberg, S.; Hoory, R. Efficient emotion recognition from speech using deep learning on spectrograms. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
46. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems, Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, NV, USA, 3–6 December 2012*; MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
47. 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]
48. Burkhardt, F.; Paeschke, A.; Rolfes, A.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
49. Lu, C.; Zheng, W.; Lian, H.; Zong, Y.; Tang, C.; Li, S.; Zhao, Y. Speech Emotion Recognition via an Attentive Time-Frequency Neural Network. *arXiv* **2022**, arXiv:2210.12430. [CrossRef]
50. Abdulmohsin, H.A.; Wahab, H.B.A.; Hossen, A.M.J.A. A new proposed statistical feature extraction method in speech emotion recognition. *Comput. Electr. Eng.* **2021**, *93*, 107172. [CrossRef]
51. Ilyosov, A.; Kutlimuratov, A.; Whangbo, T.-K. Deep-Sequence–Aware Candidate Generation for e-Learning System. *Processes* **2021**, *9*, 1454. [CrossRef]
52. 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]
53. Abdusalomov, A.B.; Mukhiddinov, M.; Kutlimuratov, A.; Whangbo, T.K. Improved Real-Time Fire Warning System Based on Advanced Technologies for Visually Impaired People. *Sensors* **2022**, *22*, 7305. [CrossRef]