electronics

MDPI

Article

Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database

Yeonguk Yu and Yoon-Joong Kim *

Department of Computer Engineering, Hanbat National University, Daejeon 34158, Korea; ryk012@naver.com
* Correspondence: yjkim@hanbat.ac.kr; Tel.: +82-42-821-1143

Received: 30 March 2020; Accepted: 23 April 2020; Published: 26 April 2020

Abstract: We propose a speech-emotion recognition (SER) model with an "attention-long Long Short-Term Memory (LSTM)-attention" component to combine IS09, a commonly used feature for SER, and mel spectrogram, and we analyze the reliability problem of the interactive emotional dyadic motion capture (IEMOCAP) database. The attention mechanism of the model focuses on emotion-related elements of the IS09 and mel spectrogram feature and the emotion-related duration from the time of the feature. Thus, the model extracts emotion information from a given speech signal. The proposed model for the baseline study achieved a weighted accuracy (WA) of 68% for the improvised dataset of IEMOCAP. However, the WA of the proposed model of the main study and modified models could not achieve more than 68% in the improvised dataset. This is because of the reliability limit of the IEMOCAP dataset. A more reliable dataset is required for a more accurate evaluation of the model's performance. Therefore, in this study, we reconstructed a more reliable dataset based on the labeling results provided by IEMOCAP. The experimental results of the model for the more reliable dataset confirmed a WA of 73%.

Keywords: speech-emotion recognition; attention mechanism; LSTM

1. Introduction

The emotional state of a person influences their modes of interactions, such as facial expressions, speech characteristics, and the content of communication. Since speech is one of the main modes of expression, a human–machine interface must recognize, understand, and respond to emotional stimuli contained in human diction. Emotions affect both vocal and verbal content. In this study, we aim to develop a mechanism that can recognize emotions from the acoustic features of utterances [1].

Several studies on speech-emotion recognition have aimed to identify features that enable the discrimination of emotions [2,3]. Various short-term and long-term features have been proposed [4], but it is still unclear which features contain more information about emotions. The most common method of emotion recognition is to extract a large number of statistical features on an utterance, reduce the dimensions using a dimension reduction technique, and classify the features using machine learning algorithms [5–7].

Recently, researchers have developed deep learning models with low-level descriptors (LLD) as inputs [1–3]. The models recognized emotions well. LLD is an acoustic feature extracted within a short frame of time, and it is mainly generated in a frame of 20–50 ms. This model mainly consists of a recurrent neural network and an attention mechanism suitable for analyzing sequential data. The commonly used LLD for emotion recognition is the IS09 feature, which is extracted using the openSMILE toolkit [8]. The IS09 feature is a 32-dimensional feature vector consisting of a fundamental frequency, the voicing probability, the frame energy, the zero-crossing rate, and 12 mel frequency cepstral coefficients (MFCC) and their first-order derivatives. It uses a 25-ms long window. IS09 is considered better than other features [1]; therefore, it has been used in various studies [9,10] on SER.

Additionally, SER attention mechanisms have been used in various fields, such as neural-machine translation [11], image-caption generation [12], and time-series prediction [13]. This mechanism generates output vectors by weighting the input vectors in the order of importance and adding them. The simplest form of an attention mechanism consists of a single query vector, $q \in \mathbb{R}^N$. In this attention mechanism, attention for a given input matrix is calculated by dot production to identify how much each vector from the matrix is similar to a query vector. The query vector, q, is used for a given input matrix $H = (h_1, h_2, \ldots, h_T)$, where $T \times N$ is the shape, T is length, and N is the dimension of the matrix H. The weight, α, for each feature vector included in the feature matrix is generated by the softmax function using the query vector as follows:

$$\alpha_t = \frac{e^{q^{\mathrm{T}} h_t}}{\sum_i^T e^{q^{\mathrm{T}} h_i}} \tag{1}$$

c, which is the final output of the attention mechanism, is generated by multiplying the weight by the input vector as follows:

$$c = \sum_i^T \alpha_i h_i \tag{2}$$

In this case, c is a context vector created by concentrating important information in the matrix by calculating the similarity between the query vector and the feature matrix. According to [1,6], this technique to generate the context vector performs better than other techniques, such as mean pooling.

In this study, we use the ability of deep learning models to extract features from the LLD and attempt to use better features when using only IS09 features. For this, we leverage the attention mechanism's ability to combine IS09 and mel spectrogram. We propose a feature-combined attention mechanism structure that integrates IS09 and mel spectrogram. The integration is then used to develop a model that predicts emotions using long short-term memory (LSTM), a dense layer, and an attention mechanism.

This attention mechanism computes the IS09-mel-spectrogram feature by selectively weighting the components of IS09 and mel spectrogram that have a significant influence on the predictive emotion. By using the feature-combined attention mechanism, we can reduce the curse of dimensionality [14] or complexity, that accompanies high-dimensional features.

Additionally, we performed an experiment to analyze the interactive emotional dyadic motion capture (IEMOCAP) [15] dataset, which is mainly used in SER research. The dataset consists of five sessions, and each session is recorded by a group of evaluators. It is composed of an improvised dataset on topics that can elicit specific emotions and a scripted dataset. For each utterance in the dataset, each evaluator judged anger, sadness, happiness, disgust, fear, surprise, frustration, excitement, and contentment, and then labeled the utterance with the most selected emotion. The label of the sample obtained through this process may be less reliable. Therefore, we reconstructed a dataset with utterances whose labels are chosen by three or more evaluators for higher reliability. A more reliable dataset such as this will help the authors derive model-evaluation results with higher accuracy. Thus, we studied the IS09–mel-spectrogram attention mechanism and the reliability of the IEMOCAP dataset.

The contributions of this paper can be summarized as follows:

- the addition of the deep-learning-based LLD feature extraction capability for enhanced and expanded IS09 features,
- an IS09-mel-spectrogram attention mechanism that focuses on the important parts, and
- an accurate-experimentation method with a highly reliable dataset.

To compare and analyze the performance of the proposed model and the IEMOCAP dataset, we used five cross-fold validations for speaker-independent evaluations and four emotions (happiness, sadness, anger, and neutral) for achieving consistency with existing research. The proposed model

achieved similar results as those of the baseline research [9]; however, with a more reliable dataset, the model achieved high performance.

The rest of the paper is organized as follows. Section 2 describes the related works. Section 3 describes the feature-combined attention model proposed. Section 4 describes the IEMOCAP dataset. Sections 5 and 6 detail three sets of experiments. The first experiment is performed to compare the performance of the model. The second one is for visualizing the attention weight of the feature-combined attention mechanism. The third experiment is for obtaining more accurate performance using a more reliable dataset. In Section 7, we discuss the problem of more reliable samples. Section 8 concludes the work.

2. Related Works

Distinguishing features are important for recognizing the speech emotion using traditional-classification techniques [16]. For example, spectrum features like MFCCs, linear frequency cepstral coefficients (LFCC), or paralinguistic features like F0 can be used [17]. In Reference [18], the authors presented a method based on the Gaussian-mixture model classifier and MFCC as features for emotion recognition. In Reference [19], the authors presented a method based on the hidden Markov model and support vector machine (SVM) for emotion recognition using MFCC and LFCC. In Reference [20], the authors proposed the combined feature with MFCC and the residual-phase feature for music-emotion recognition using autoassociative neural networks, support vector machines, and radial-basis function neural networks.

With numerous successful applications of deep neural networks (DNNs), more researchers began to focus on emotion-recognition DNNs. In Reference [21], authors proposed a generalized discriminant analysis based on DNNs to learn discriminative features of low dimensions optimized with respect to a fast classification from a large set of acoustic features for emotion recognition. They show a highly significant improvement over the SVM. In Reference [22], authors used CNNs for face and CNNs for voice to recognize the emotion of a given video. They proposed a strong model for facial-expression emotion recognition, achieving state-of-the-art performance. In Reference [23], the authors proposed a deep dual recurrent encoder model that utilizes text data and audio signals simultaneously to obtain a better understanding of speech data. Their model encodes the information from text and audio sequences using RNNs. In Reference [24], the authors proposed a model consisting of CNN and RNN that automatically learns the best representation of the speech signal directly from the raw-time representation. In Reference [10], the authors used the bidirectional long- and short-term memory (BLSTM) model to capture important information from the speech signal.

Also, with numerous successful applications of the attention mechanism, more researchers began to focus on DNNs with attention mechanisms. In Reference [25], the authors proposed a three-dimensional attention-based convolutional recurrent neural network to learn discriminative features for SER, where the mel spectrogram with deltas and delta–deltas were used as inputs. They assumed that calculating the deltas and delta–deltas for personalized features preserve the effective emotional information and reduce the influence of emotional, irrelevant factors. In Reference [26], the authors proposed a deep recurrent neural network model with the attention mechanism for SER. Their model is based on the intuition that it is beneficial to emphasize the expressive part of the speech signal for emotion recognition. In Reference [27], the authors used an architecture involving both convolutional layers, for extracting high-level features from raw spectrograms, and recurrent ones for aggregating long-term dependencies. In Reference [28], the authors used the RNN model for extracting statistical functionals over speech segments. In Reference [29], the authors proposed an attentive convolutional neural network consisting of a CNN layer, a max-pooling layer, and an attention layer. In Reference [9], the authors used the BLSTM model with an attention layer. This research shows the effectiveness of the attention mechanism for SER.

Our works differ from the works mentioned above. Our attention-LSTM-attention model uses two attention layers. The first attention layer is a layer using an IS09-mel-spectrogram-combined attention

mechanism that focuses on the important parts for the input of the model, and the second attention layer is a layer using a temporal attention mechanism for emphasizing the temporally important parts of a given audio signal by weighting the LSTM output. To the best of our knowledge, this is the first attempt to use an attention mechanism for combining many features to use the input of the SER model.

3. Model

3.1. Problem Definition

The emotion recognition task is as follows. The input of the model is an n-dimensional feature sequence data of length T, $\boldsymbol{X} = (X_1, X_2, \ldots, X_T) \in \mathbb{R}^{N \times T}$. Therefore, $X_t = (x_1, x_2, \ldots, x_N)$ is an N-dimensional feature vector at time t, and $X^n = \left(x^1, x^2, \ldots, x^T\right)$ is the nth feature vector of size T. The purpose for the model is to take $\boldsymbol{X}$ as the input and predict the label $\overline{y}$. The real labels for $\boldsymbol{X}$, y are anger, neutral, sadness, and happiness. Finally, the model can be defined by Equation (3) and is shown in Figure 1.

$$\overline{y} = F(\boldsymbol{X}) \tag{3}$$

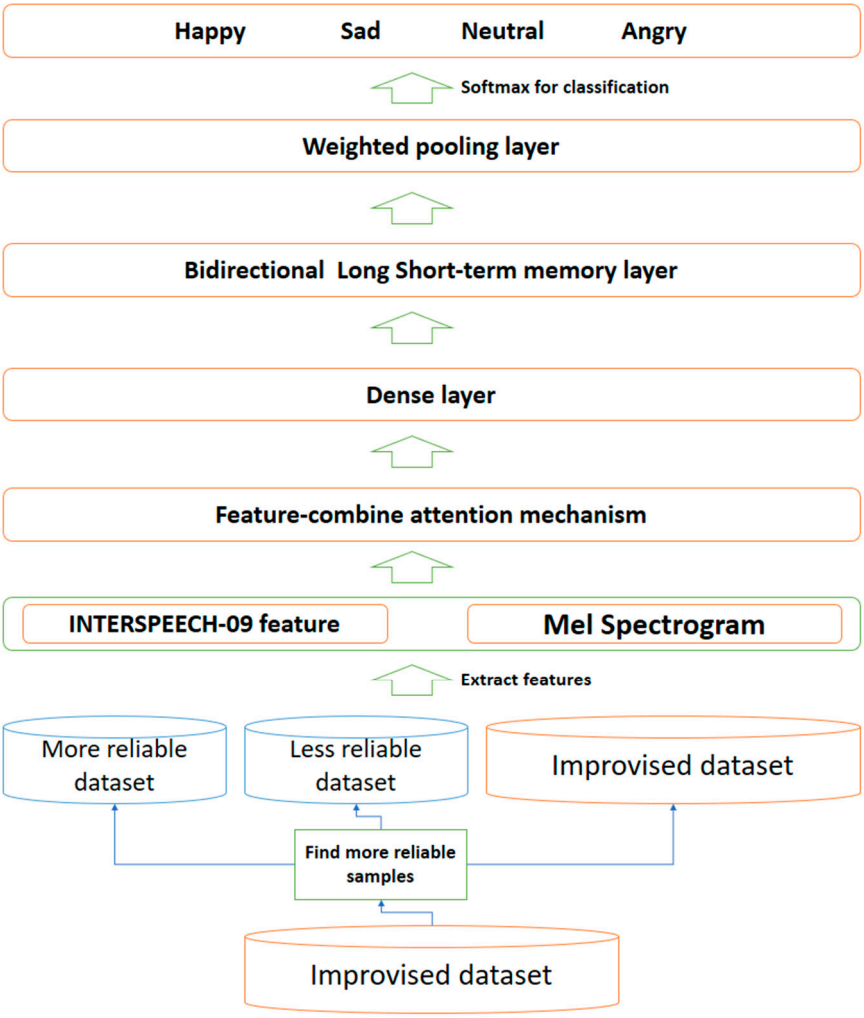

Figure 1. Architecture of our proposed research.

3.2. Feature-Combined Attention Mechanism

The feature-combined attention mechanism weighs on both the feature axis and the time axis, whereas the conventional-attention mechanism weighs on the time axis. This is to focus on the important feature axis for combining features. A representation of the mechanism is shown in Figure 2.

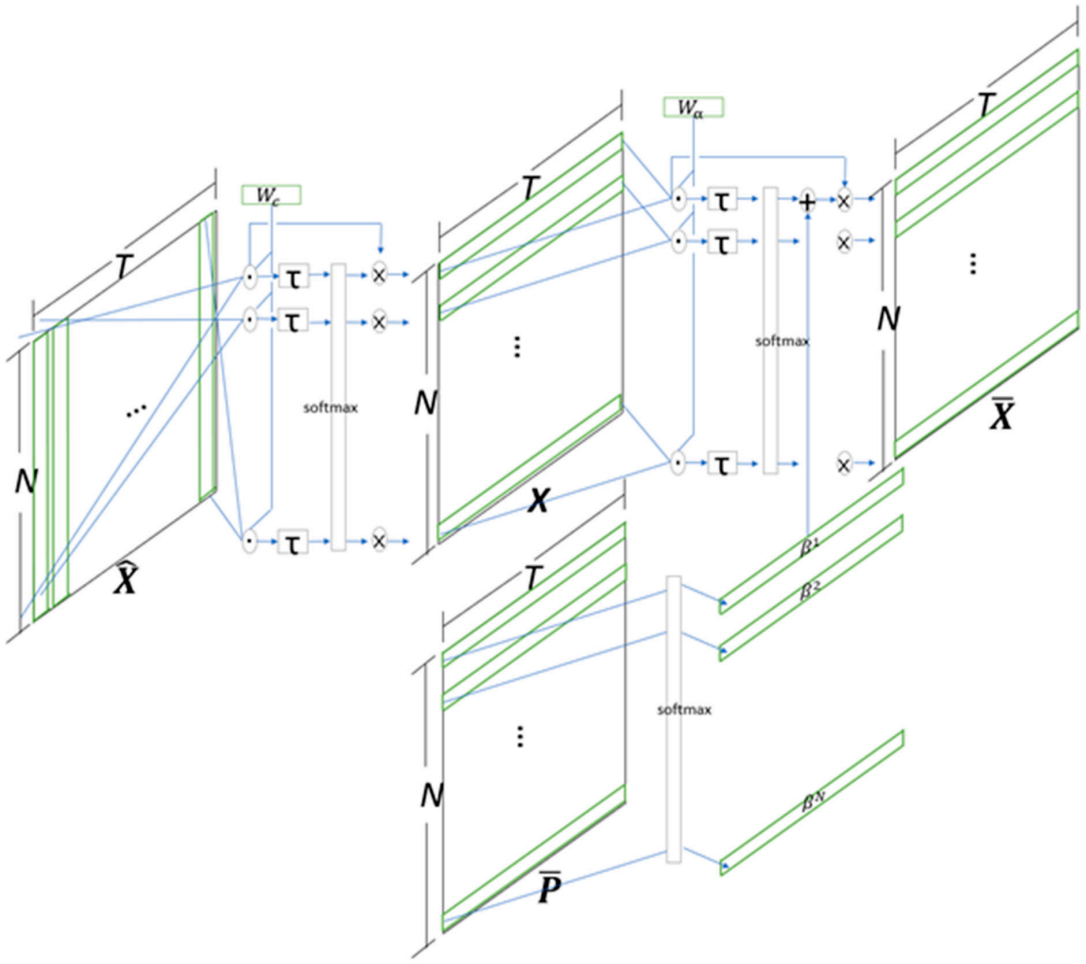

Figure 2. Representation of the attention mechanism for feature selection (·, τ, +, × refer to dot production, tanh, plus, and elementwise multiplication, respectively).

The purpose of the attention mechanism is to generate new features by combining two input features, the IS09 and mel spectrogram. The computation of the feature-combined attention mechanism consists of three weight calculations.

The first calculation is made on a time base to ignore unimportant time. Therefore, the feature vector obtained from the mute segment of the audio can be ignored when weighting the feature axis. For this, the weight c^n for the t-th feature on the time axis, $X_t = (x_1, x_2, \ldots, x_N)$, is calculated by Equation (4), where W_c is the weight parameter used in the feed-forward network. The weights are each multiplied by the input feature vector $\hat{\boldsymbol{X}}$ to produce a feature vector $\boldsymbol{X}$ with a suppressed mute segment.

$$c^n = \frac{e^{\tan h(W_c^T X_t)}}{\sum_i^N e^{\tan h(W_c^T X_i)}} \tag{4}$$

The second calculation gives weights according to the variation of features. For each N-th feature, $X^n = \left(x^1, x^2, \ldots, x^T\right) \in \mathbb{R}^T$, $\alpha^i \in \mathbb{R}$ is calculated by Equation (5).

$$\alpha^n = \frac{e^{\tan h(W_\alpha^{\mathrm{T}} X^n)}}{\sum_i^N e^{\tan h(W_\alpha^{\mathrm{T}} X^i)}} \tag{5}$$

where W_α is a weight parameter used in the feed-forward network.

The third weight calculation uses a weight parameter, $\boldsymbol{P} = (P_1, P_2, \ldots, P_T) \in \mathbb{R}^{N \times T}$, to render weights according to the position of the feature. $\boldsymbol{P}$ passes through a dense layer with the tanh function as the active function, and it is transformed into the $\overline{\boldsymbol{P}} = \left(\overline{P}_1, \overline{P}_2, \ldots, \overline{P}_T\right) = \left(\overline{P}^1, \overline{P}^2, \ldots, \overline{P}^N\right) \in \mathbb{R}^{N \times T}$. For each $\overline{P}^n = \left(\overline{p}_1^n, \overline{p}_2^n, \ldots, \overline{p}_T^n\right) \in \mathbb{R}^T$, weight $\beta^n = \left(\beta_1^n, \beta_2^n, \ldots, \beta_T^n\right) \in \mathbb{R}^T$ is calculated by Equation (6).

$$\beta^n = \left(\frac{\overline{p}_1^n}{\sum_i^N \overline{p}_1^i}, \frac{\overline{p}_2^n}{\sum_i^N \overline{p}_2^i}, \ldots, \frac{\overline{p}_T^n}{\sum_i^N \overline{p}_T^i} \right) \tag{6}$$

After the weights, the final output $\overline{X}^n$ is calculated using the α^n and β, as in Equation (7).

$$\overline{X}^n = \left(\left(\alpha^n + \beta_1^n\right) x_1^n, \left(\alpha^n + \beta_2^n\right) x_2^n, \ldots, \left(\alpha^n + \beta_T^n\right) x_T^n \right) \tag{7}$$

Therefore, $\overline{X}^n$ is the result of the feature-combined attention mechanism created by absolute weights according to the positions of the feature and weights, according to the variation and weights, and according to the temporal importance.

4. IEMOCAP Database

The dataset used to compare the performance of the model is the improvised dataset included in IEMOCAP (Figure 3). The improvised dataset is part of the whole dataset. Unlike the scripted dataset that contains a script for speakers to read, the improvised dataset is a collection of statements for improvisation. This dataset consists of five sessions, each of which further contains samples of speeches from two speakers. We divided the audio signals into four emotional categories, happiness, sadness, neutral, and anger, for consistency with existing studies. To evaluate the performance of the emotion-recognition model independent of the speaker, the model was trained in four sessions and evaluated in the last session. Moreover, audio signals 5 s or longer were used from 0 to 5 s, and audio signals less than 5 s long were zero-padded to set a length of 5 s.

The features extracted from the audio signal are IS09 and mel spectrogram. Mel spectrogram is an 80-dimensional feature vector extracted by shifting a 20-ms-long window by 10 ms. Therefore, both feature vectors are generated from a segment at the rate of 100 frames/second (fps). As a result, a total of 500 frames of feature vectors were generated, and they were used separately or as a concatenate of 112-dimensional feature vectors.

The IEMOCAP dataset contains less reliable samples, i.e., samples in which the choices of the evaluator are staggered, and the emotions are not obvious. In order to determine whether this less reliable sample is adversely affecting the model, less reliable and more reliable samples are separated from the dataset to construct a new dataset, and are used for training and evaluation. In the same vein, we analyzed the impact of less reliable samples on the model. Information on the newly generated datasets is listed in Table 1. We define a less reliable sample as a sample in which the judgments are less than two, because the D, judgment of the evaluator is mostly inconsistent, and a more reliable sample as a sample in which D is mostly consistent. Therefore, a more reliable sample may be considered to have a correct label, a clear characteristic of the emotion included in the sample. Therefore, the model will perform better with more reliable samples for training and evaluation.

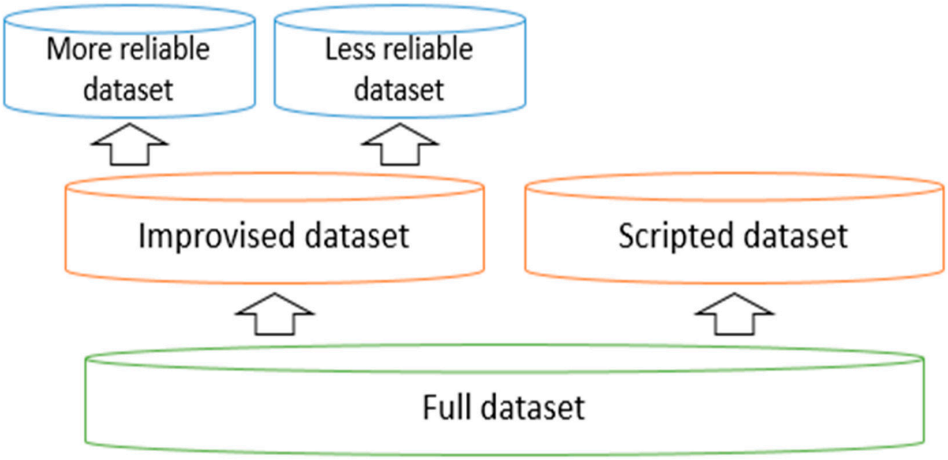

Figure 3. Interactive emotional dyadic motion capture (IEMOCAP) dataset structure.

Table 1. Specifications of the improvised dataset, more reliable dataset, and less reliable dataset.

Dataset	Number of Data Points	Happy	Angry	Neutral	Sad	Decision
I (improvised)	2280	284	289	1099	608	-
R (more reliable)	1202	99	132	582	389	$D > 2$
U (less reliable)	1078	185	157	517	219	$D <= 2$

5. Experiment

The model used in the experiment uses the feature-integrated attention mechanism on top of the existing model. The model in (1) uses rectified linear unit (ReLU) dense layers and BLSTM recurrent layers. Weighted pooling by an attention mechanism is used for producing the final output. We named this model the "LSTM-attention model (LA)". Similarly, the model that included the feature-combined attention mechanism in front of the dense layer of the LA model is coined as the "attention-LSTM-attention model (ALA)". We used three types of features: IS09, mel spectrogram, and IS09 + mel spectrogram. Therefore, the experiments are classified in Table 2.

Table 2. Specification of each experiment.

Experiment Name	Used Feature	Model
IS_LA	IS09	LA
MS_LA	MelSpectrogram	LA
ISMS_LA	IS09+MelSpectrogram	LA
IS_ALA	IS09	ALA
MS_ALA	MelSpectrogram	ALA
ISMS_ALA	IS09+MelSpectrogram	ALA

In each experiment, five sessions were trained and evaluated with five cross-validation folds, and then overall accuracy (weighted accuracy, WA) and average recall over the different emotional categories (unweighted accuracy, UA) were averaged and used as the performance metrics of the model as described in [1]. We implemented the model on TensorFlow [30] and trained via the Adam optimizer set at 80 epochs with a 128-size minibatch at a learning rate of 0.001. To overcome the imbalance of label categories, we weighted the loss function of each category c with $W_c = \frac{N_{total}}{N_c}$, where N_c is the number of samples from category c, and N_{total} is the total number of samples.

The experimental results in Table 3 reveal that the ALA model using IS09 and mel spectrogram shows the best performance. This result appears to improve the performance of the deep learning

model by reducing the complexity by the feature-combined attention mechanism, focusing on the important parts of the various features only. However, when the IS09 was used as a feature, the ALA model did not perform well, which may be the result of dividing the IS09 feature into important and unnecessary parts despite it being made only for SER without the unnecessary parts. Unlike IS09, mel spectrogram, where important and unimportant parts coexist, performs better.

Table 3. Experimental results in the form of mean accuracies ± standard deviations.

Experiment	Weighed Accuracy	Unweighted Accuracy
IS_LA	60.99 ± 2.8	60.41 ± 2.3
MS_LA	65.21 ± 3.1	63.27 ± 2.1
ISMS_LA	64.27 ± 4.7	61.82 ± 5.7
IS_ALA	55.74 ± 3.1	53.86 ± 2.2
MS_ALA	66.07 ± 2.6	63.20 ± 1.5
ISMS_ALA	67.66 ± 3.4	65.08 ± 4.5

In addition, the visualization of the attention weight for the examination of the behavior of the feature-combined attention mechanism, we confirmed that the attention mechanism actually gave less weight to the features that were not necessary for the recognition of emotion and more weight to the necessary ones. The visualization results are shown in Figure 4, which shows the weight of the attention when the feature vectors are a concatenation of IS09 and mel spectrogram. From the figure, the weights are applied differently to mel spectrogram and IS09 according to the input.

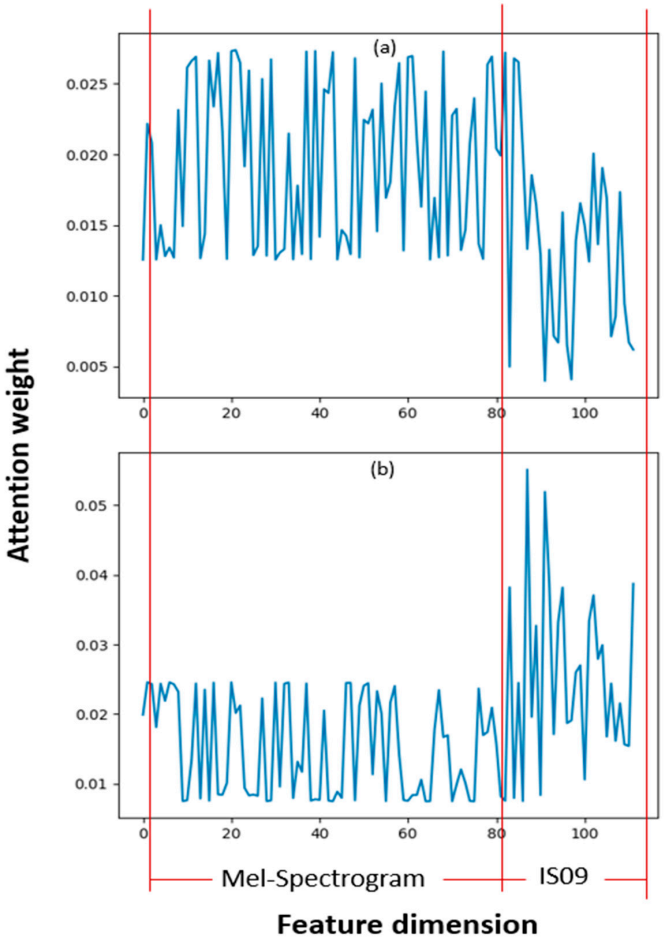

Figure 4. Feature-selected attention-weight visualization. (**a**) Mel spectrogram feature, as weighted by the attention mechanism. (**b**) IS09 feature, as weighted by the attention mechanism.

6. Analysis of IEMOCAP Database

Table 4 compares the performance of the proposed method with some state-of-the-art methods proposed in other studies. They conducted all their experiments using the improvised dataset of the IEMOCAP dataset.

From Table 4, the model proposed in this study shows almost the same performance as the model proposed in [9]. Changing the parameters several times over and altering the environment of the experiment also shows similar results, which is considered a limitation of the dataset. The problem is the reliability of the IEMOCAP dataset. In order to confirm this, we experimented with the improvised (I), more reliable (R), and less reliable (U) datasets described in Section 4 and trained and evaluated the model for each dataset.

The experiment was performed using the ALA model. First, as shown in Table 1 of Section 4, I, R, and U datasets were used for training and evaluation. For the training, training data of I (TI), R (TR), and U (TU) were used. For the evaluation, evaluation data of I (EI), R (ER), and U (EU) were used. The parameter setting and the environment were the same as in the previous experiment. The results of the experiment are shown in Table 5.

Table 4. Comparison of performance with other studies.

Experiment	Weighed Accuracy	Unweighted Accuracy
Etienne et al. [28]	64.5	61.7
Tzinis et al. [29]	64.2	60.0
Lee et al. [10]	62.9	63.9
Neumann et al. [30]	62.1	_
Ramet et al. [9]	68.8	63.7
ISMS_ALA (proposed model)	67.66 ± 3.4	65.08 ± 4.5

Table 5. Results of the ALA model for improvised (I), more reliable (R), and less reliable (U) datasets.

Training Dataset	Evaluation Dataset	WA	UA
TI	EI	67.66 ± 3.4	65.08 ± 4.5
	ER	73.20 ± 7.4	68.37 ± 7.2
	EU	62.60 ± 4.4	61.96 ± 6.4
TR	EI	65.68 ± 3.5	60.31 ± 5.4
	ER	73.18 ± 7.7	69.43 ± 7.0
	EU	59.58 ± 5.4	57.17 ± 4.5
TU	EI	62.63 ± 4.4	60.69 ± 5.8
	ER	69.22 ± 7.0	66.22 ± 5.7
	EU	60.70 ± 4.5	61.57 ± 7.0

The experimental results reveal that the model trained with TI performed better than the models trained with TR and TU in the EI and EU tests since I contains R and U. Therefore, the model can be optimized for EI and EU. All models exhibited their best performances for ER, confirming the high accuracy of the labels of the more reliable data. Therefore, ER should be evaluated for accurate performance comparison. The proposed model shows a higher performance of WA (73.2%) and UA (68.37%).

7. Discussion

However, it is not confirmed that the more reliable data can be seen as more reliable for evaluation. Because more reliable samples can be seen as just easier samples, which are probably positioned in a farther distance from emotional boundaries in the feature space. In this paper, we believed comparing the evaluation on more reliable datasets is a more accurate comparison. Because, if the more reliable samples are positioned in a farther distance from other reliable samples from other emotional categories,

classification models would be better at mapping categories for inputs in the feature space. Even so, we believe it is required to do more research for a more reliable sample that classification models recognized as wrong emotions to find the reason.

8. Conclusions

In this study, we attempted to improve the performance of the SER model by combining the IS09 features, which are mainly used in SER and mel spectrogram, an LLD, and using them as inputs. For this purpose, after the concatenation of IS09 and mel spectrogram, the attention mechanism was used to combine the features by weighting to the appropriate part. Since then, the same model with the dense layer and bidirectional LSTM has been used. The model was tested using the IEMOCAP dataset. Experimental results show that the ALA model for SER improved by approximately 3% over the LA model in terms of weighted accuracy and unweighted accuracy.

However, it was also determined that the performance improvement was not a reliability issue stemming from the IEMOCAP dataset labeling method. We isolated the reliability problem of the IEMOCAP dataset through additional experiments. For this, we believed the evaluation for the dataset with samples that two or more people agreed on the emotion can show accurate results. This model presented a WA of 73% and a UA of 68%.

We note that more reliable datasets could be tested more for checking the reliability of the dataset. Therefore, we could conduct experiments to check the reliability of the dataset by using the clustering technique or dimension-reduction technique, and clarify how far more reliable samples and less reliable samples are away from the class boundaries for future work. Furthermore, we could conduct experiments to compare the proposed model with other models.

Author Contributions: Formal analysis, Y.Y.; investigation, Y.Y.; methodology, Y.-J.K.; project administration, Y.-J.K.; software, Y.Y.; writing—original draft, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. 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.
2. Tahon, M.; Devillers, L. Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 16–28. [CrossRef]
3. Schuller, A.B.; Batliner, D.; Seppi, S.; Steidi, T.; Vogt, J.; Wagner, L.; Devillers, L.; Vidrascu, N.; Kessous, A.L. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In Proceedings of the 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2016; pp. 2253–2256.
4. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. *Int. J. Speech Technol.* **2012**, *15*, 99–117. [CrossRef]
5. Schuller, B.; Arsic, D.; Wallhoff, F.; Rigoll, G. Emotion recognition in the noise applying large acoustic feature sets. *Proc. Speech Prosody* **2006**, *2006*, 276–289.
6. Álvarez, A.; Cearreta, I.; López-Gil, J.-M.; Arruti, A.; Lazkano, E.; Sierra, B.; Garay-Vitoria, N. Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in Spoken Spanish and Standard Basque Language. In Proceedings of the International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 11–15 September 2006; Volume 4188, pp. 565–572.
7. Busso, C.; Bulut, M.; Narayanan, S.S. *Toward Effective Automatic Recognition Systems of Emotion in Speech in Social Emotions in Nature and Artifact: Emotions in Human and Human-Computer Interaction Social Emotions in Nature and Artifact: Emotions in Human and Human-Computer Interaction*; Gratch, J., Marsella, S., Eds.; Oxford University Press: New York, NY, USA, 2013; pp. 110–127.

8. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia—MM '13, Barcelona, Spain, 21–25 October 2013; pp. 835–838. [CrossRef]
9. Ramet, G.; Garner, P.N.; Baeriswyl, M.; Lazaridis, A. Context-Aware Attention Mechanism for Speech Emotion Recognition. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 126–131. [CrossRef]
10. Lee, J.; Tashev, I. High-level feature representation using recurrent neural network for speech emotion recognition. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 1537–1540.
11. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
12. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
13. Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G.W. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2627–2633.
14. Verleysen, M.; Francois, D. The Curse of Dimensionality in Data Mining and Time Series Prediction. In Proceedings of the International Work-Conference on Artificial Neural Networks, Barselona, Spain, 8–10 June 2005; Volume 3512, pp. 758–770.
15. 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. Eval.* **2008**, *42*, 335–359. [CrossRef]
16. 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]
17. Vidrascu, L.; Devillers, L. Five emotion classes detection in real-world call centre data: The use of various types of paralinguisticfeatures. In Proceedings of the International Workshop on Paralinguistic Speech Between Models and Data, Saarbrücken, Germany, 2–3 August 2007; DFKI Publications: Kaiserslautern, Germany, 2007.
18. Kandali, C.A.B.; Routray, A.; Basu, T.K. Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In Proceedings of the 2008 IEEE Region 10 Conference (TENCON 2008), Hyderabad, India, 19–21 November 2008; pp. 1–5.
19. Chenchah, F.; Lachiri, Z. Acoustic Emotion Recognition Using Linear and Nonlinear Cepstral Coefficients. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*, 135–138. [CrossRef]
20. Nalini, N.J.; Palanivel, S. Music emotion recognition: The combined evidence of MFCC and residual phase. *Egypt. Inform. J.* **2016**, *17*, 1–10. [CrossRef]
21. Stuhlsatz, A.; Meyer, C.; Eyben, F.; Zielke, T.; Meier, G.; Schuller, B. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 5688–5691.
22. Albanie, S.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference—MM '18, Seoul, Korea, 22–26 October 2018; pp. 292–301.
23. Yoon, S.; Byun, S.; Jung, K. Multimodal Speech Emotion Recognition Using Audio and Text. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118. [CrossRef]
24. 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; pp. 5200–5204.
25. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]

26. Hsiao, P.-W.; Chen, C.-P. Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2526–2530.
27. Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L.; Schmauch, B. Speech emotion recognition with data augmentation and layer-wise learning rate adjustment. *arXiv* **2018**, arXiv:1802.0563*68*.
28. Tzinis, E.; Potamianos, A. Segment-based speech emotion recognition using recurrent neural networks. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 190–195.
29. Neumann, M.; Vu, N.T. Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. In Proceedings of the Interspeech 2017—18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 1263–1267. [CrossRef]
30. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. *OSDI* **2016**, *16*, 265–283.