symmetry

MDPI

Article

Multi-Task Conformer With Multi-Feature Combination for Speech Emotion Recognition

Jiyoung Seo and Bowon Lee *

Department of Electrical and Computer Engineering, Inha University, Incheon 22212, Korea; seo.jiyoung@dsp.inha.ac.kr
* Correspondence: bowon.lee@inha.ac.kr; Tel.: +82-32-860-7423

Abstract: Along with automatic speech recognition, many researchers have been actively studying speech emotion recognition, since emotion information is as crucial as the textual information for effective interactions. Emotion can be divided into categorical emotion and dimensional emotion. Although categorical emotion is widely used, dimensional emotion, typically represented as arousal and valence, can provide more detailed information on the emotional states. Therefore, in this paper, we propose a Conformer-based model for arousal and valence recognition. Our model uses Conformer as an encoder, a fully connected layer as a decoder, and statistical pooling layers as a connector. In addition, we adopted multi-task learning and multi-feature combination, which showed a remarkable performance for speech emotion recognition and time-series analysis, respectively. The proposed model achieves a state-of-the-art recognition accuracy of 70.0 ± 1.5 % for arousal in terms of unweighted accuracy on the IEMOCAP dataset.

Keywords: speech emotion recognition; arousal; valence; spoken language understanding

Citation: Seo, J.; Lee, B. Multi-Task Conformer With Multi-Feature Combination for Speech Emotion Recognition. *Symmetry* **2022**, *14*, 1428. https://doi.org/10.3390/sym14071428

Academic Editor: Giuseppe Bagliesi

Received: 13 June 2022
Accepted: 10 July 2022
Published: 12 July 2022

1. Introduction

Speech is one of the richest sources of information for interactions such as context, emotion, and speaker's identification. In recent years, automatic speech recognition (ASR) has significantly advanced, achieving a single-digit word error rate (WER) percentage even without applying a language model [1,2]. In addition to ASR, speech emotion recognition (SER) is also important because speech may carry a speaker's emotional state, which may provide information as significant as the textual information, for effective communication. For example, there exist an anger detector [3] for sensing disgruntled customers and a depression detector [4] for medical screenings.

In the SER field, emotion can be interpreted as categorical emotion or dimensional emotion. Categorical emotion indicates emotions such as anger, happiness, and sadness. Dimensional emotion represents emotions pertaining to arousal (or activation), valence, and dominance on individual axes [5]. Categorical emotion is more intuitive and commonly used than dimensional emotion, which may require additional interpretation. As such, studies on categorical emotion have been conducted actively compared with dimensional emotion. On the contrary, dimensional emotion can represent human emotion in a wider range than that of categorical emotion. In addition, distinguishing categorical emotion tends to cause confusion if arousal and valence levels are similar [6,7]. Therefore, in this study, we focus on the arousal and valence of dimensional emotion, and, in particular, on discrete arousal and valence tasks, since arousal and valence recognition can be designed as a regression task [8–10] or as a categorical task [11–13].

To extract the essential representation of the speech feature, it is important for SER models to focus on both local and global characteristics [1,14]. Conformer [1] is a model devised to simultaneously represent local characteristics through a convolutional neural network (CNN) and global characteristics through Transformer [15]. In recent ASR studies,

Conformer-based models have shown significant performance improvements [1,16,17]. Furthermore, a Conformer-based model has shown a remarkable performance for categorical SER in a recent study [18]. Due to the lack of existing work on using Conformer-based models for dimensional SER, we adopted the Conformer network for dimensional emotion recognition in terms of arousal and valence in this study.

Multi-task learning (MTL) refers to learning multiple relevant tasks at the same time, and has an advantage in preventing a model from overfitting to a single task [12]. It has been widely adopted for SER studies [12,13,19], dividing multiple tasks into a primary task and auxiliary tasks and using weighted-sum-loss as the objective function. The main contribution of this study is the adoption of the Conformer model for dimensional SER. In particular, we reduced the number of parameters of the original Conformer model and applied MTL to be suitable for the SER task because SER generally has fewer labels than ASR. An additional contribution is the introduction of feature stacking [20] for the SER task to further improve the accuracy. The proposed multi-task Conformer-based SER (MTC-SER) model achieved a state-of-the-art (SOTA) performance in arousal recognition in terms of unweighted accuracy (UA) on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [5].

2. Related Works

In this section, we describe recent studies on SER, especially the recognition of discrete arousal and valence, the Conformer, and several topics relevant to SER.

2.1. Deep Learning-Based Speech Emotion Recognition

In recent years, with the emergence of deep neural networks, many SER models have adopted deep learning networks, including Transformer-based models for categorical emotion [21,22]; basic deep learning structures such as a CNN, recurrent neural network, and dense network for regression-based dimensional emotion recognition [9,10]; and adversarial autoencoders (AAE) [23] for discrete dimension emotion [13,24]. Unfortunately, SER has a limitation, which is the lack of labeled emotion datasets [13]. Therefore, many studies have attempted to solve the problem of limited labeled datasets. For instance, semi-supervised learning with adversarial autoencoders [13,24] and transfer learning with ASR datasets [25] have been proposed. Many studies [13,21,24] have typically adopted the cross-validation strategy, which divides a dataset into k smaller subsets, called folds. The folds for validation and testing are chosen k times, and the k performance results are averaged. However, there exist subtle differences in the number of folds: 5-fold [21] and 10-fold [11,13,24,26] and whether the fold is speaker-dependent [24] or not [11,26], or both [13]. In this study, we selected the semi-supervised adversarial autoencoder model from [24] as the previous SOTA model since it has the exactly same speaker-dependent 10-fold strategy for fair comparison. We also compare our model with [13] using the same AAE model as [24] with speaker-independent 5-fold for discrete dimensional SER.

2.2. Conformer

Conformer was devised to combine the advantages of Transformer [15] and convolution capturing the global contexts and local contexts, respectively. It has achieved significant performance improvements and has become the base model for several ASR models, with exceptionally low WER [1,16,17]. Conformer has a sequence of modules as a block. The conformer block comprises a feed forward module, multi-head self-attention module, convolution module, and another feed forward module.

A feed forward module is composed of a normalization layer, two linear layers, one activation function, and a dropout. The first linear layer expands the dimension by a expansion factor. Then, the second linear layer decreases the dimensions to the original value. A swish function [27] is used as an activation function. The multi-head self attention module has a structure that places multi-head attention between the normalization layer and the dropout. The convolution module contains three convolution layers; that is, two

pointwise convolution layers and a 1D depthwise convolution layer; a normalization layer, activation functions, and batch normalization. The first pointwise convolution layer expands the channel by an expansion factor. All modules have a residual unit connecting the module input to the module output without going through the module. An overview of the Conformer structure is depicted in Figure 1.

In [1], the authors proposed three types of models—Conformer S, Conformer M, and Conformer L—that consist of a Conformer encoder and recurrent neural networks (RNNs) as the decoder. The models have the same structure but a different number of hyperparameters to vary the model sizes from 10.3 M parameters to 118.8 M parameters. The hyperparameters are the number of encoder layers, which is the number of Conformer blocks, encoder dimension, attention heads, convolution kernel size, decoder layers, and decoder dimension. However, the convolution kernel size and number of decoder layers are consistently set to 32 and 1, respectively. The detailed values of the hyperparameters for each of the three models are shown in Table 1. For all models, the number of attention heads, convolution kernel size, and the number of decoder layers are set to 4, 32, and 1, respectively.

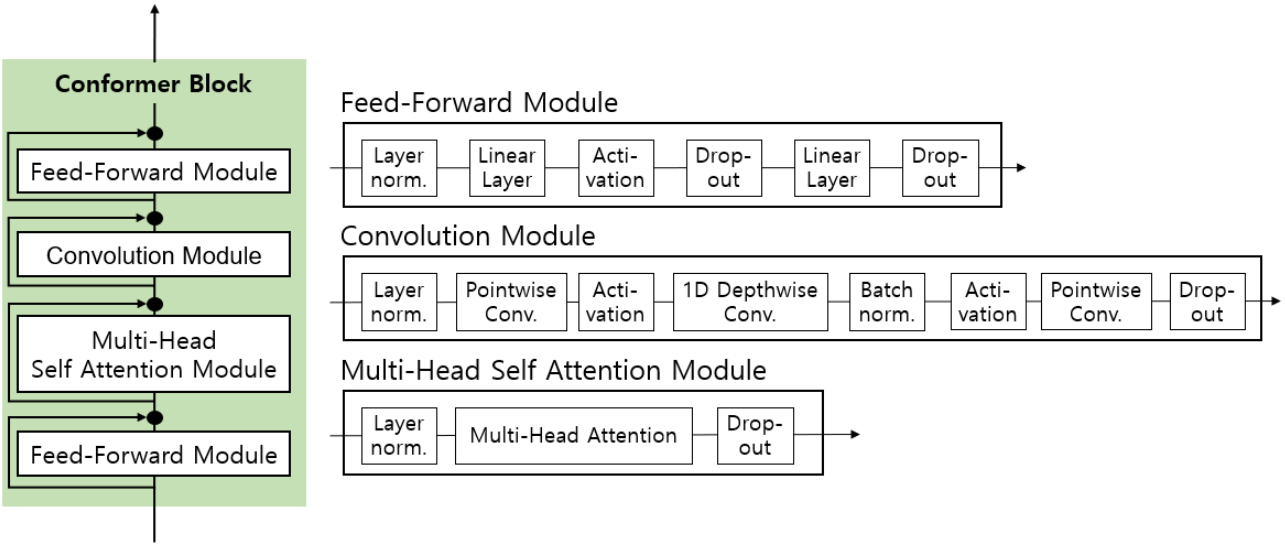

Figure 1. Overview of the Conformer block structure and block modules.

Table 1. Hyperparameters of Conformer models [1].

Model	Conformer S	Conformer M	Conformer L
Number of Parameters	10.3×10^6	30.7×10^6	118.8×10^6
Encoder Layers	16	16	17
Encoder Dimension	144	256	512
Decoder Dimension	320	640	640

2.3. Multi-Task Learning

Multi-task learning has shown successful improvement in SER [12,13,19]. This mitigates the overfitting problem [12] caused by small datasets and training with a single task. It categorizes multiple tasks into two groups, the primary task for the main purpose and auxiliary tasks for the overfitting problem. The loss for each task is summed with different weights:

$$L_T(\mathbf{W}) = \sum_N \alpha_n L_n(\mathbf{W}), \tag{1}$$

where $L_T(\cdot)$ is a loss function with network weights $\mathbf{W}$, which is a summation of each loss for task n, $L_n(\cdot)$, with a loss weight α_n.

2.4. Speech Features

A feature study is one of the main topics for SER research [28,29]. Therefore, many features for SER tasks are studied and used, such as the short-time Fourier transform (STFT) or spectrogram [13,21], Mel-spectrogram (Mel) [30,31], and Mel-frequency cepstral coefficients (MFCCs) [31,32]. The STFT obtained by applying the fast Fourier transform during short time intervals represents the frequency change over time [33]. Mel rescales the STFT with the Mel-frequency scale modeling of human hearing [34]. Then, the MFCCs are obtained by conducting a ceptral analysis of the Mel to acquire useful features to predict the formant frequencies [35]. As STFT, Mel, and MFCCs are suitable for analyzing speech information, these features are generally used for SER.

3. Proposed Method

We designed our MTC-SER model in consideration of three aspects: input feature, structure, and loss function. Multi-feature stacking was applied for the input feature, and the Conformer structure, which was designed for the ASR task, was redesigned for the SER task. Finally, multi-task learning was adopted to boost the performance. The detailed model structure consisting of these three aspects is explained in this section.

3.1. Multi-Feature Combination

Features are an important topic for SER [28,29]. Conventionally, a large amount of research on SER uses only a single feature as the input [36]. However, a recent study on appliance classification showed stacking multiple features extracted from sequential data can improve the performance [20]. Thus, we conducted a multi-feature combination (MFC) experiment with three general features, namely STFT, Mel, and MFCCs. In addition, to analyze the temporal variations in the audio data, we added the velocity (ΔMFCCs) and acceleration (ΔΔMFCCs) for the MFCCs. Finally, the MFC experiment was divided into three stacking types: a single feature with STFT (FS1); a triple-feature with STFT, Mel, and MFCCs (FS3); and the MFC of all (FS5), as shown in Figure 2. The input feature can be represented as $X = \{x_1, x_2, ..., x_L\}$, where $x_i \in \mathbb{R}^{d_{feat}}$. L is the temporal dimension of the spectrogram, and d_{feat} denotes the dimension of stacked features, which is calculated as a sum of the individual features' dimensions.

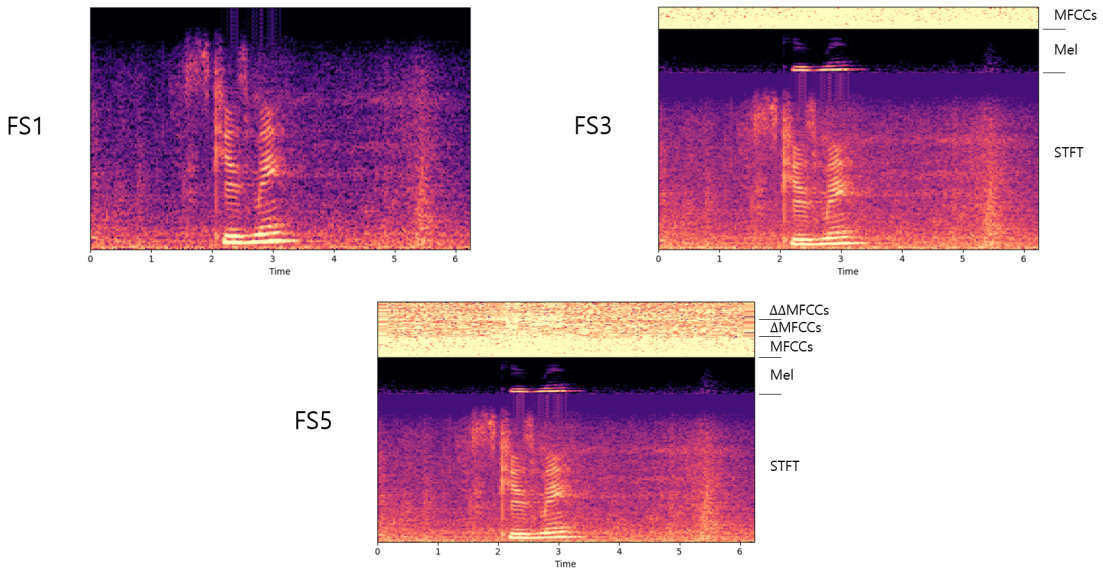

Figure 2. Examples of FS1, FS3, and FS5.

3.2. Conformer-Based Model

Our model consists of a Conformer encoder, statistical pooling layers, and fully connected layers. The Conformer encoder extracts a meaningful hidden vector from a given input. Then, the statistical pooling layer compresses the hidden vector into a fixed-length hidden vector, and the fully connected layers link the fixed-length hidden vector to each task. The connection between all layers and modules is described in Figure 3, and details of the model structure are provided below.

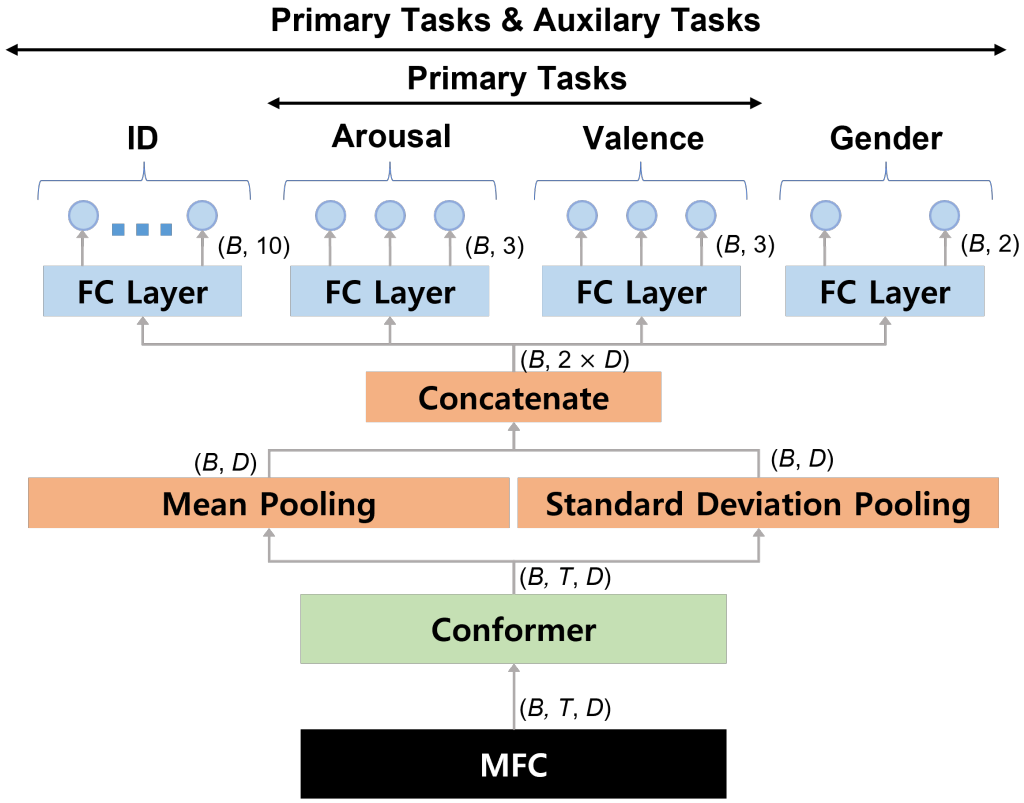

Figure 3. MTC-SER structure, where B is the batch size, T is the longest temporal length in the batch, and D is the hidden dimension.

In this study, we adopted the Conformer encoder for two reasons. First, our model needs to take features extracted from the speech signal, such as ASR tasks, and second, the SER model needs to capture not only the feature relationships within a short time span but also speech features in a wider time span, such as intonations, to more accurately predict arousal and valence. In addition, we set our model with the same encoder setting as Conformer M and reduced the parameter size to 17.4 million, which is approximately half the original size of 31.4 million in Table 1. This is because our model has fewer classes than the original Conformer. The Conformer encoder can be represented as follows:

$$H_{encoder} = \text{Conformer}(X) \tag{2}$$

where $H_{encoder} \in \mathbb{R}^{d_{Enc} \times L}$, and d_{Enc} indicates the encoder dimension. After the Conformer encoder, the statistical pooling layers first calculate the mean and standard deviation along the frame dimension, and then concatenate the results.

Statistical pooling, such as mean pooling and standard derivation pooling, is widely used for condensing features with variable frame length to fixed-length features [37–39]. In the proposed model, statistical pooling layers convert the Conformer encoder's output vector in variable sizes depending on the input speech length into a fixed-length vector such that the layers connect the Conformer encoder to the fully connected (FC) layers. The statistical pooling layers calculate the output as follows:

$$H_{mean} = \text{mean}(H_{encoder}) \tag{3}$$

$$H_{std} = \text{std}(H_{encoder}) \tag{4}$$

$$H_{pooled} = \text{concatenate}(H_{mean}, H_{std}) \tag{5}$$

where $H_{mean}, H_{std} \in \mathbb{R}^{d_{Enc}}$, and $H_{pooled} \in \mathbb{R}^{2d_{Enc}}$.

Our model yields more than one output for several tasks because multi-task learning is applied, which will be discussed in Section 3.3. All tasks needs to share a model but have a different number of classes. Therefore, the FC layers decode a common hidden vector for each task as follows:

$$Y_{Arousal} = \text{FCLayer}(H_{pooled}) \tag{6}$$

$$Y_{Valence} = \text{FCLayer}(H_{pooled}) \tag{7}$$

$$Y_{ID} = \text{FCLayer}(H_{pooled}) \tag{8}$$

$$Y_{Gender} = \text{FCLayer}(H_{pooled}) \tag{9}$$

where $Y_{Arousal} \in \mathbb{R}^{d_{Arousal}}$, $Y_{Valence} \in \mathbb{R}^{d_{Valence}}$, $Y_{ID} \in \mathbb{R}^{d_{ID}}$, and $Y_{Gender} \in \mathbb{R}^{d_{Gender}}$. $d_{Arousal}$, $d_{Valence}$, d_{ID}, and d_{Gender} are the number of classes for arousal, valence, ID, and gender, respectively.

3.3. Multi-Task Learning

For the four tasks, we categorized arousal and valence recognition as two primary tasks and the classification of ID and gender as auxiliary tasks. Two types of MTL experiments were conducted: primary MTL and primary and auxiliary MTL using the weighted-sum loss defined in Equation (1):

$$L_P(\mathbf{W}) = \alpha_{Arousal} L_{Arousal}(\mathbf{W}) + \alpha_{Valence} L_{Valence}(\mathbf{W}) \tag{10}$$

$$\begin{aligned} L_{AP}(\mathbf{W}) = \alpha_{Arousal} L_{Arousal}(\mathbf{W}) + \alpha_{Valence} L_{Valence}(\mathbf{W}) \\ + \alpha_{ID} L_{ID}(\mathbf{W}) + \alpha_{Gender} L_{Gender}(\mathbf{W}) \end{aligned} \tag{11}$$

where $L_P(\cdot)$ and $L_{AP}(\cdot)$ are the loss functions for the primary MTL and primary and auxiliary MTL, respectively.

4. Experimental Setup

In this section, we define the experimental setup for the proposed model and features. Then, the strategy for the comparative experiments and the results are presented.

4.1. Model

Our model shown in Figure 3 consists of a Conformer-based encoder, statistical pooling layers as a connector, and FC layers as a decoder. Each FC layer for a single task is constructed with an input size of 512 and an output size equivalent to the number of classes for each classification task. We made significant changes to the original Conformer model to make it appropriate for the SER task. We reduced the size of the original Conformer encoder because our model needs fewer classes compared with typical ASR models. We used eight Conformer blocks (or encoder layers) with hidden dimension D of 256, four attention-heads, and a convolution kernel size of 31 for symmetric padding. We used an expansion factor of 2 for the convolution module and an expansion factor of 4 for the feed forward module, and the dropout rate was 0.1, equivalent to the original Conformer.

4.2. Dataset

We used IEMOCAP [5] to evaluate the performance of MTC-SER, which is a multi-modal and multi-speaker database and contains approximately 12 hours of audio and video recordings of 5 female and 5 male actors. The data are labeled with emotion, annotated

into categorical and dimensional emotions, along with transcripts, speaker ID, gender, and motion capture of the face. The categorical emotions include seven emotions, along with other and neutral, and the dimensional emotions include activation, valence, and dominance, ranging from 1 to 5. In accordance with the previous SOTA study [24], we converted values from 1 to 5 into three levels by mapping $[1, 2]$ to low/negative, $(2, 3.5]$ to medium/neutral, and $(3.5, 5]$ to high/positive for arousal/valence labels.

4.3. Feature Extraction and MFC

We extracted all of the features after downsampling the audio data from 48 kHz to 16 kHz sampling rate and applying a hamming window of size 400 (25 ms) and hop size of 160 (10 ms) using the Librosa [40]. STFT, Mel, and MFCCs were represented with 512 frequency bins, 64 filters, and 32 filters, respectively. All of the features were scaled in decibels and normalized before stacking. We stacked the features vertically in three different combinations: FS1, FS3, and FS5, as shown in Figure 2. Then, a dimension of input becomes the sum of feature dimensions. For example, in the case of FS3, d_{feat} is $353(257 + 64 + 32)$. In addition, to prevent overfitting and to augment the data, we applied SpecAugmentation [41] with the parameters listed in Table 2.

Table 2. Parameters of SpecAugmentation. According to [41], SpecAugmentation is conducted with a time warp parameter W, frequency mask parameter F, and time mask parameter T. The maskings were applied m_F and m_T times for frequency masking and time masking, respectively. p is the upper bound of the time mask.

W	F	m_F	T	p	m_T
0	15	3	45	1.0	3

4.4. Training and Evaluation

In multi-task leaning, it is important to set the loss weights [12,13,19]. We conducted two sets of experiments on MTL. First, we compared two cases of MTL, one being primary MTL and the other primary and auxiliary MTL. Primary MTL refers to multi-task learning with only two primary tasks: arousal and valence recognition; therefore, the weights of each loss, $\alpha_{Arousal}$ and $\alpha_{Valence}$, were set to 0.5. Primary and auxiliary MTL refers to multi-task learning with two primary tasks and two auxiliary tasks (speaker ID and gender recognition). The weights for the two primary tasks, $\alpha_{Arousal}$ and $\alpha_{Valence}$, were set to 0.45, and the weights for the two auxiliary tasks, α_{ID} and α_{Gender}, were set to 0.05 to make the model focus on the primary tasks.

After comparing the two MTL types, we compared three different loss weights strategies: major, neutral, and minor for the primary and auxiliary MTL that showed an improved performance. Major puts heavier weights on the primary tasks with the same loss weight used for the primary and auxiliary MTL as in the previous experiment. In the case of neutral, all loss weights are equivalently set to 0.25. Minor puts less weight on the primary tasks by setting the loss weights to 0.05, 0.05, 0.45, and 0.45 for arousal, valence, speaker ID, and gender, respectively. Further, we used the Adam optimizer with a learning rate of 1×10^{-5} and a weight decay of 5×10^{-5} to optimize the model.

We adopted 10-fold cross-validation to train and evaluate MTC-SER and stratify each fold to be based on speakers as in the previous SOTA study [24] to provide speaker information for auxiliary task labeling. To evaluate our models, we selected the models showing the best arousal or valence accuracy on the validation set for each experimented model. We used unweighted accuracy (UA) as the evaluation metric and calculated the mean and standard deviation of all UAs for each fold.

5. Results

The proposed MTC-SER, which is the best model among the combinations from Section 3.3, outperforms the previous SOTA model [24] using adversarial autoencoders,

as shown in Table 3, with a UA of 70.0 ± 1.5 %. In addition, it shows that our model still outperforms the same AAE model [13] with a speaker-independent 5-fold strategy. However, the valence recognition of MTC-SER showed a lower accuracy than the previous SOTA model.

Table 3. Comparison with the previous models.

Model	Arousal	Valence
Semi-supervised AAE (2020) [13]	64.5 ± 1.5	62.2 ± 1.0
Semi-supervised AAE (2019) [24]	64.81	**64.77**
MTC-SER	**70.0 ± 1.5**	60.8 ± 1.3

To compare the performance dependency on the two MTL types, the primary MTL and the primary and auxiliary MTL, and three MFC types, namely, FS1, FS3, and FS5, we conducted experiments for all combinations. The results are shown in Table 4. With all MFC types, the primary and auxiliary MTL shows higher UAs of arousal and valence than the primary MTL. In addition, MFC showed an increased performance in arousal recognition. For FS3 and FS5, the models could record higher arousal accuracies than models with a single feature. However, it can be seen that valence accuracies are not significantly affected by MFC.

Table 4. Experimental results of MTL types.

	MTL Type			
	Primary		Primary and Auxiliary	
MFC Type	Arousal	Valence	Arousal	Valence
FS1	68.1 ± 2.5	60.0 ± 1.7	68.7 ± 2.2	60.0 ± 1.3
FS3	68.2 ± 1.7	60.0 ± 1.8	**69.0 ± 2.1**	**60.7 ± 2.0**
FS5	68.3 ± 2.3	58.8 ± 1.5	69.0 ± 2.8	59.7 ± 1.8

The primary and auxiliary MTL showed better arousal UAs than the primary MTL and the best valence UA in Table 4. Therefore, we compared the three loss weight strategies of the primary and auxiliary MTL type with all MFC types. The results for arousal and valence are shown in Tables 5 and 6, respectively. It can be seen that MFC improved arousal accuracies but did not affect the valence accuracies, as shown in Table 4. Additionally, the best loss weights strategies for arousal and valence are different. Arousal accuracies are improved when loss weights of primary tasks are equal to or less than auxiliary tasks. In contrast, valence accuracies tend to improve when loss weights of primary tasks are higher than auxiliary tasks.

Table 5. Arousal results depending on loss weight strategies for primary and auxiliary MTL.

Loss Weight Strategy		Major	Neutral	Minor
MFC Type	FS1	68.7 ± 2.2	**69.2 ± 3.1**	68.7 ± 1.6
	FS3	69.0 ± 2.1	**69.7 ± 1.7**	69.1 ± 1.9
	FS5	69.0 ± 2.8	69.9 ± 2.0	**70.0 ± 1.5**

Table 6. Valence results depending on loss weight strategies for primary and auxiliary MTL.

Loss Weight Strategy		Major	Neutral	Minor
MFC Type	FS1	**60.0 ± 1.3**	59.4 ± 2.1	52.9 ± 1.8
	FS3	60.7 ± 2.0	**60.8 ± 1.3**	53.1 ± 1.7
	FS5	**59.7 ± 1.8**	59.0 ± 2.0	59.0 ± 2.0

6. Discussion

In this study, MTC-SER was proposed for dimensional speech emotion recognition. MTC-SER with the primary and auxiliary tasks showed better UAs for arousal and valence than with the primary tasks. However, the best arousal and valence recognition accuracies were obtained with different loss weights strategies. It can be considered that, if the model fits better to a specific task and a loss of the task is relatively lower than other tasks, the model is trained better with a small loss weight for that task so that the model can focus more on other less fitted tasks.

MTC-SER even without MFC showed a higher arousal accuracy than the previous SOTA model. Furthermore, the results show that a combination of multiple unimodal features as an input is effective for arousal recognition. Still, our model showed lower valence recognition accuracies than the previous SOTA model, and it was shown that the valence recognition accuracies were less affected by MFC. It can be regarded that statistical pooling layers, the connector in MTC-SER, may not be adequate to carry contextual information for valence recognition.

7. Conclusions

In this study, we proposed MTC-SER, consisting of Conformer, MTL, and MFC. The proposed Conformer-based model achieved superior arousal recognition accuracies for all sets of experiments compared with the previous SOTA model. In addition to Conformer, the application of the auxiliary MTL and MFC further improved the arousal recognition accuracy. Therefore, we can conclude that adopting the Conformer encoder is effective for arousal recognition and has synergies with auxiliary multi-task learning and multi-feature stacking.

For valence, we plan to study the design of better pooling layers and apply the output of ASR as another auxiliary task for incorporating more contextual information. In addition, a comparison with other datasets will be studied as future work.

Author Contributions: Conceptualization, J.S. and B.L.; methodology, J.S.; software, J.S.; validation, J.S.; formal analysis, J.S.; investigation, J.S.; resources, B.L.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, B.L.; visualization, J.S.; supervision, J.S.; project administration, J.S.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A01087325), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University) and RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: This study is conducted with the IEMOCAP dataset. IEMOCAP is available upon request from https://sail.usc.edu/iemocap/ accessed on 9 June 2022.

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

References

1. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. In Proceedings of the INTERSPEECH, ISCA, Shanghai, China, 25–29 October 2020; Volume 2020, pp. 5036–5040.
2. Xu, Q.; Baevski, A.; Likhomanenko, T.; Tomasello, P.; Conneau, A.; Collobert, R.; Synnaeve, G.; Auli, M. Self-training and pre-training are complementary for speech recognition. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Toronto, ON, Canada, 6–11 June 2021; pp. 3030–3034.
3. Burkhardt, F.; Ajmera, J.; Englert, R.; Stegmann, J.; Burleson, W. Detecting anger in automated voice portal dialogs. In Proceedings of the INTERSPEECH, ISCA, Pittsburgh, PA, USA, 17–21 September 2006; Volume 2006, pp. 1053–1056.

4. Huang, Z.; Epps, J.; Joachim, D. Speech landmark bigrams for depression detection from naturalistic smartphone speech. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, UK, 12–17 May 2019; pp. 5856–5860.
5. 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]
6. Singh, P.; Saha, G.; Sahidullah, M. Deep scattering network for speech emotion recognition. *arXiv* **2021**, arXiv:2105.04806.
7. Provost, E.M.; Shangguan, Y.; Busso, C. UMEME: University of Michigan emotional McGurk effect data set. *IEEE Trans. Affect. Comput.* **2015**, *6*, 395–409. [CrossRef]
8. Parthasarathy, S.; Busso, C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In *Interspeech*; ISCA: Stockholm, Sweden, 2017; Volume 2017, pp. 1103–1107.
9. Chen, J.M.; Chang, P.C.; Liang, K.W. Speech Emotion Recognition Based on Joint Self-Assessment Manikins and Emotion Labels. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), IEEE, San Diego, CA, USA, 9–11 December 2019; pp. 327–3273.
10. Atmaja, B.T.; Akagi, M. Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information. In Proceedings of the 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, Yangon, Myanmar, 5–7 November 2020; pp. 166–171.
11. Metallinou, A.; Wollmer, M.; Katsamanis, A.; Eyben, F.; Schuller, B.; Narayanan, S. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affect. Comput.* **2012**, *3*, 184–198. [CrossRef]
12. Zhang, Z.; Wu, B.; Schuller, B. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Brighton, UK, 12–17 May 2019; pp. 6705–6709.
13. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J.; Schuller, B.W. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Trans. Affect. Comput.* **2020**, *11*, 992–1004. [CrossRef]
14. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 5998–6008.
16. Zhang, Y.; Qin, J.; Park, D.S.; Han, W.; Chiu, C.C.; Pang, R.; Le, Q.V.; Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv* **2020**, arXiv:2010.10504.
17. Chan, W.; Park, D.; Lee, C.; Zhang, Y.; Le, Q.; Norouzi, M. SpeechStew: Simply mix all available speech recognition data to train one large neural network. *arXiv* **2021**. arXiv:2104.02133.
18. Shor, J.; Jansen, A.; Han, W.; Park, D.; Zhang, Y. Universal Paralinguistic Speech Representations Using Self-Supervised Conformers. *arXiv* **2021**, arXiv:2110.04621.
19. Xia, R.; Liu, Y. A multi-task learning framework for emotion recognition using 2D continuous space. *IEEE Trans. Affect. Comput.* **2017**, *8*, 3–14. [CrossRef]
20. Kim, J.G.; Lee, B. Appliance classification by power signal analysis based on multi-feature combination multi-layer LSTM. *Energies* **2019**, *12*, 2804. [CrossRef]
21. Wang, X.; Wang, M.; Qi, W.; Su, W.; Wang, X.; Zhou, H. A Novel end-to-end Speech Emotion Recognition Network with Stacked Transformer Layers. In Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Toronto, ON, Canada, 6–11 June 2021; pp. 6289–6293.
22. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Interspeech*; ISCA: Graz, Austria, 2019; pp. 2803–2807.
23. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. *arXiv* **2015**, arXiv:1511.05644.
24. Rana, R.; Latif, S.; Khalifa, S.; Jurdak, R.; Epps, J. Multi-task semisupervised adversarial autoencoding for speech emotion. *arXiv* **2019**, arXiv:1907.06078.
25. Tits, N.; Haddad, K.E.; Dutoit, T. Asr-based features for emotion recognition: A transfer learning approach. *arXiv* **2018**, arXiv:1805.09197.
26. Wu, J.; Dang, T.; Sethu, V.; Ambikairajah, E. A Novel Markovian Framework for Integrating Absolute and Relative Ordinal Emotion Information. *arXiv* **2021**, arXiv:2108.04605.
27. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for activation functions. *arXiv* **2017**, arXiv:1710.05941.
28. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
29. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*; ISCA: Singapore, 2014; Volume 2014, pp. 223–227.
30. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]

31. Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; Tarokh, V. Speech emotion recognition with dual-sequence LSTM architecture. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Barcelona, Spain, 4–8 May 2020; pp. 6474–6478.
32. 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]
33. Allen, J.B.; Rabiner, L.B. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE* **1977**, *65*, 1558–1564. [CrossRef]
34. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
35. Logan, B. Mel frequency cepstral coefficients for music modeling. In Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR), Plymouth, MA, USA, 23–25 October 2000.
36. Singh, Y.B.; Goel, S. A systematic literature review of speech emotion recognition approaches. *Neurocomputing* **2022**, *492*, 245–263. [CrossRef]
37. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
38. Lozano-Diez, A.; Plchot, O.; Matejka, P.; Gonzalez-Rodriguez, J. DNN based embeddings for language recognition. In Proceedings of the In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, AB, Canada, 15–20 April 2018; pp. 5184–5188.
39. Cooper, E.; Lai, C.I.; Yasuda, Y.; Fang, F.; Wang, X.; Chen, N.; Yamagishi, J. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Barcelona, Spain, 4–8 May 2020; pp. 6184–6188.
40. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
41. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech* **2019**, *2019*, 2613–2617. [CrossRef]