electronics

MDPI

Article

Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation

John Lorenzo Bautista [1,2], Yun Kyung Lee [3] and Hyun Soon Shin [1,2,3,*]

[1] Artificial Intelligence Department, University of Science and Technology Korea, Daejeon 34113, Republic of Korea
[2] Emotion (Brain-Emotion) Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea
[3] Emotion Information Communication Technology Industrial Association, Daejeon 34111, Republic of Korea
* Correspondence: hsshin@etri.re.kr

Abstract: In this paper, an automatic speech emotion recognition (SER) task of classifying eight different emotions was experimented using parallel based networks trained using the Ryeson Audio-Visual Dataset of Speech and Song (RAVDESS) dataset. A combination of a CNN-based network and attention-based networks, running in parallel, was used to model both spatial features and temporal feature representations. Multiple Augmentation techniques using Additive White Gaussian Noise (AWGN), SpecAugment, Room Impulse Response (RIR), and Tanh Distortion techniques were used to augment the training data to further generalize the model representation. Raw audio data were transformed into Mel-Spectrograms as the model's input. Using CNN's proven capability in image classification and spatial feature representations, the spectrograms were treated as an image with the height and width represented by the spectrogram's time and frequency scales. Temporal feature representations were represented by attention-based models Transformer, and BLSTM-Attention modules. Proposed architectures of the parallel CNN-based networks running along with Transformer and BLSTM-Attention modules were compared with standalone CNN architectures and attention-based networks, as well as with hybrid architectures with CNN layers wrapped in time-distributed wrappers stacked on attention-based networks. In these experiments, the highest accuracy of 89.33% for a Parallel CNN-Transformer network and 85.67% for a Parallel CNN-BLSTM-Attention Network were achieved on a 10% hold-out test set from the dataset. These networks showed promising results based on their accuracies, while keeping significantly less training parameters compared with non-parallel hybrid models.

Keywords: speech emotion recognition; parallel networks; attention-based network; audio data augmentation; transformer; deep learning

Citation: Bautista, J.L.; Lee, Y.K.; Shin, H.S. Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation. *Electronics* **2022**, *11*, 3935. https://doi.org/10.3390/electronics11233935

Academic Editors: Ruifeng Xu and Chiman Kwan

Received: 20 October 2022
Accepted: 24 November 2022
Published: 28 November 2022

1. Introduction

Human communication relies heavily on emotional cues and is one key aspect in improving human–computer interactions (HCI) [1]. In line with the increasing trend and continuing technological development of a Metaverse, enhanced social interactions are becoming more of a challenge with the need for sound and speech recognition, as well as emotional recognition to achieve a natural interaction and increased immersion [2]. A person's emotion influences their vocal characteristics and their linguistic contents, and thus, with the help the ever-improving computational powers of modern computers, studies on Speech Emotion Recognition (SER) systems have continuously grown with the rise in deep neural networks by mapping audio signals into feature maps representing a speech sample's vocal characteristics [3]. SER, on a machine learning perspective, is a classification problem using audio samples as input that are then classified into a set of pre-defined emotions. Emotional audio datasets are essential to the development and

evaluation of such SER systems. Labeled audio signals are used to train an SER to recognize different emotional classes. Because of this, a large number of labeled audio data are essential in developing a robust SER system.

The impressive progress in computer vision has helped researchers to improve SER by considering an audio sample's spectral features as an image input. Convolutional Neural Networks (CNN) are considered as a gold standard in image processing. This architecture consists of feature representations derived from the weights of multiple convolutional layers [4,5]. This technology can be utilized in SERs using mel-spectrograms to transform audio data into visual audio signals based on its frequency components. These image-like representations can then be trained on a CNN network as if they were images. However, traditional CNN networks take only a single frame of input and do not perform computations on a timestep sequence, which means that they cannot remember past data from the same sample when processing the next timestamp.

In this study, our goal is to identify underlying emotions in speech using voice-based feature extraction methods, by utilizing the power of CNNs while addressing the issue of the small number of available training data using different data augmentation techniques. To improve on the CNN's current architecture, we applied a CNN network parallel to time distributed models of LSTMs with attention-based models, and a transformer-based architecture to capture spatial information and features, while helping the network learn and predict frequency distributions of emotions over time. The spectrograms were processed as images, with the height and width being represented by the time and frequency scales of the spectrogram, using CNN's established strengths in image classification and spatial feature representations, while temporal feature representations were represented by attention-based models Transformer and BLSTM-Attention modules.

Two main models (Parallel CNN-BLSTM-Attention, and Parallel CNN-Transformer models) are proposed in this paper. These network models were then trained on the Ryeson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6] dataset, with some of their audio signals augmented to increase the training samples using multi-fold audio data augmentation techniques such as Additive White Gaussian Noise (AGWN), SpecAugment, Room Impulse Response (RIR), and Tanh Distortion.

2. Related Works

2.1. Speech Emotion Recognition

Emotion recognition from speech has been studied to an extent as it plays an important role in improving human–computer interactions. Speech Emotion Recognition (SER) has been developed as a system that can identify the multiple emotional states from different audio samples. Traditionally, emotion recognition has been developed using classical machine learning techniques such as Hidden Markov Models (HMM) [7], Gaussian Mixture Models (GMM) [8,9], Support Vector Machines (SVM) [10], and k-nearest Neighborhood Classifiers (kNN) [9,11]. In recent years, deep learning-based classifiers have become the common approach to SER systems such as Deep Neural Networks (DNN) [12,13], Deep Boltzmann Machine (DBM) [14], Convolutional Neural Network (CNN) [15], Recurrent Neural Networks (RNN) [16], and Long Short-Term Memory (LSTM) [17,18].

2.2. Ryeson Audio-Visual Database of Emotional Speech and Song

The Ryeson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6] was utilized in this study. In the RAVDESS dataset, 12 actors and 12 actresses each performed eight different emotions, namely neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. These emotions are performed twice, in two different forms, by singing and speaking sentences for each emotion. However, for this study only recordings that were performed by speaking are used. Table 1 shows the speech sample count for each of the said emotions. Each speech sample has a length of 4 s, each with a 1 s silence at the beginning and at the end of each recording.

Table 1. RAVDESS Dataset Speech Samples.

Emotion	Speech Samples
Neutral	96
Calm	192
Happy	192
Sad	192
Angry	192
Fearful	192
Disgust	192
Surprised	192
Total	1440

2.3. Neural Network Approach for SER Based on RAVDESS

There were neural network-based implementations of SER for the RAVDESS dataset in recent years. Zeng et al. [19] implemented a deep neural network based on spectrograms extracted from songs and speech utterances from the RAVDESS dataset. The spectrogram inputs were used on a multi-task gated residual network and achieved an accuracy of 65.97% on the test data. Similarly, Popova et al. [20] used a spectrogram extracted only from speech utterances from the dataset and used a VGG-16 convolutional neural network to achieve an accuracy of 71% on the test data. To improve on the models of such CNN based SER systems, Issa et al. [21] used a deep convolutional neural network (DCNN) to perform the SER task. By utilizing combined transformed audio signals of different methods such as MFCC, Chromagram, Mel-spectrogram, Spectral contrast, and Tonnetz representation as a concatenated input for their network, the researchers obtained a 71.61% accuracy on all eight emotional classes. However, one disadvantage of these models is that latent space weights of a CNN model represent and focus on the spatial properties of the audio signals. Since speech depends on sequences over time, spatial features should also be considered. To address this, Li et al. [22] experimented using the speech utterances from the dataset and implemented a multimodal deep learning approach to perform a fine-grained emotion recognition which uses temporal alignment mechanism to capture fine-grain emotions. They obtained an accuracy of 64% on a combined CNN and LSTM network, and 66.5% on an LSTM with Attention Network using acoustic information alone, and 70.8% whilst using semantic embeddings. In recent years, pre-trained models such as Wav2Vec encoders were introduced as a self-supervised approach, which was found to be helpful for the speech emotion recognition tasks [23–25]. This method for speech emotion recognition utilizes an end-to-end model using raw signals, which is a pre-trained model using large-scale voice data with their encoder used as a feature extraction model. Although end-to-end based deep learning models have advantages on calculating features, pre-training on large scale data is usually time consuming and is more complex for the initial model weight values. Application of hybrid models in SER has been adapted in the past. There were previous papers that proposed the use of LSTM and Transformer networks with CNN [26]. Han et al. [27] utilized a network of ResNet18 combined with CNN and Transformer in a parallel architecture using MFCC features on the RADVESS dataset. Recently, Slimi et al. [27] applied a hybrid time-distributed CNN-Transformer for the SER task, which reported an accuracy of 82.72% on 8:1:1 hold out test. These architectures maximized the benefits of CNN's capability of learning from small quantities of data and Transformer's superior learning capabilities. However, such architectures could be more resource intensive as they have huge trainable parameters compared to ensemble networks running in parallel.

2.4. Improving Networks Using Data Augmentation

One problem for training neural networks is the limited number of data available for training and evaluation. Networks that are trained on a very small number of data are at risk of over-fitting. When networks are over-fit, their task to classify other unseen datasets

could not be met and could generally impair their robustness. To tackle this problem data augmentations are used to artificially add replicas of the training data, whilst preserving the labels from the dataset. This technique has been used in other audio-based classification tasks. Data augmentation methods can be classified into traditional methods which augment on raw audio signal, and augmentation on the spectral representation [28,29]. Audio augmentation techniques have also been used on SER systems and have been shown to improve the models' robustness and classification accuracies [30]. A summary of the reviewed literature of SER systems trained on the RAVDESS dataset is shown in Table 2.

Table 2. Comparison of reviewed literatures of SER system trained on RAVDESS dataset.

Year	Related Works	Input Features	Accuracy (%)
2017	VGG-16, Popova et al. [20]	Mel-Spectrogram	71.00
2020	Multi-gated Residual Network, Zeng. et.al [19]	Mel-Spectrogram	65.97
2020	DCNN, Issa et al. [21]	MFCC, Chromagram, Mel-spectrogram, Spectral contrast, and Tonnetz representation	71.61
2020	Fine Grained Model with Temporal Alignment, Li et.al [22]	Multimodal	64.00 (CNN + LSTM) 66.50 (LSTM + Attention) 70.80 (with Semantic Embeddings)
2020	Wav2Vec 2.0 embeddings, Pepino et al. [24]	Raw Audio	84.30 (Pre-trained) 68.70 (Fined-tined)
2021	Resnet Transformer-Encoder CNN Han et al. [26]	MFCC	80.89
2022	Time Distributed CNN-Transformer Slimi et al. [27]	Mel-Spectrogram	82.13 (TDCNN + Vision Transformer) 82.72 (TDCNN + Transformer)

3. Proposed Work

This section contains two main parts: (A) Parallel CNN-Based Classification Models, and (B) Data augmentation.

3.1. Parallel CNN-Based Classification Models

CNN based models are widely used in image processing tasks and have proven their power in capturing spatial features from speech by considering a spectrogram as a single grayscale image with the time features as the width and frequencies as the height. Attention-based models, on the other hand, are widely used in speech, video, and other tasks requiring temporal features. Two main attention-based models are focused on this paper: LSTM with Attention (BLSTM-Attention-CNN), and Transformer-Encoder (Transformer-CNN). An LSTM with attention is a model composed of a bidirectional LSTM as an encoder and decoder that utilizes attention weights within sequences, with the idea of freeing the encoder-decoder architecture from the fixed-length internal representation. The Transformer network, on the other hand, is designed for the network to learn to predict frequency distributions of different classes according to the global structure of the spectrogram of each training emotion. In contrast with LSTM, transformers would not only learn to predict variations according to time steps, but also look at multiple previous time steps through their multi-head self-attention layers. The model architectures of the parallel models are shown in Figure 1.

The main motivation of this paper is to be able to use both the power of a CNN network to capture spatial features, and attention-based networks LSTM-Attention and Transformer to capture the temporal features using a spectrogram in a speech emotion recognition task by training on an expanded RAVDESS dataset using only simple augmentation techniques, and compared to a VGG-16 model as its baseline along with other CNN based architectures.

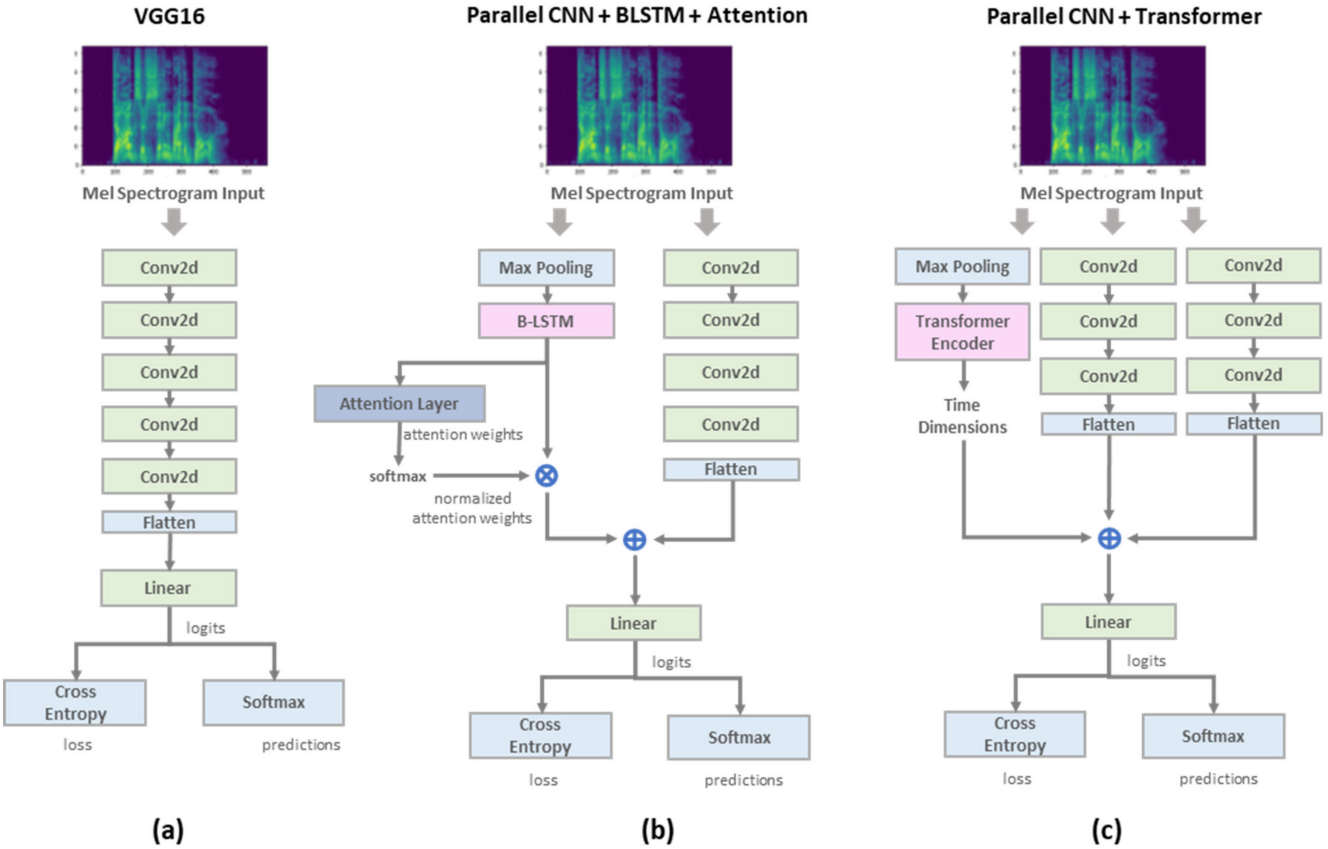

Figure 1. Model Architectures for VGG-16 as a baseline (**a**), proposed BLSTM-Attention-CNN network (**b**), and proposed Transformer-CNN network (**c**).

3.2. Data Augmentation Techniques

Data augmentation is a technique used for improving the performance of most machine learning models by artificially increasing data used in training. Because of the small size of the dataset, the model becomes prone to overfitting. This problem could be eased by generating more samples for the training data. In this paper, we used different data augmentation techniques such as Additive White Gaussian Noise (AWGN), SpecAugment [1], Room Impulse Response (RIR) based augmentation, and Tanh Distortion to augment the RAVDESS dataset's training data. Although some of these data augmentation techniques such as AWGN and RIR are already considered as data augmentation standards, we introduce a multi-fold augmentation practice on the experiments by applying these augmentation techniques which have shown increased accuracy during evaluation.

3.2.1. Additive White Gaussian Noise

In AWGN, a gaussian noise vector sampled from a normal distribution with a zero-mean time average is added uniformly across the frequency distribution. Implementation of noise addition only requires the summation of two signals, and the signal-to-noise ratio (SNR) of the output can be manipulated through scaling the signal. This SNR is randomized and selected uniformly in the decibel scale, which fits a more logarithmic scale rather than linear, similar to the human hearing. The use of Additive White Noise has shown positive impact on the accuracy of different speech and audio-based classification tasks [31]. A comparison of the original audio signal's waveform and mel-spectrogram compared with the AWGN augmented signal is shown on Figure 2.

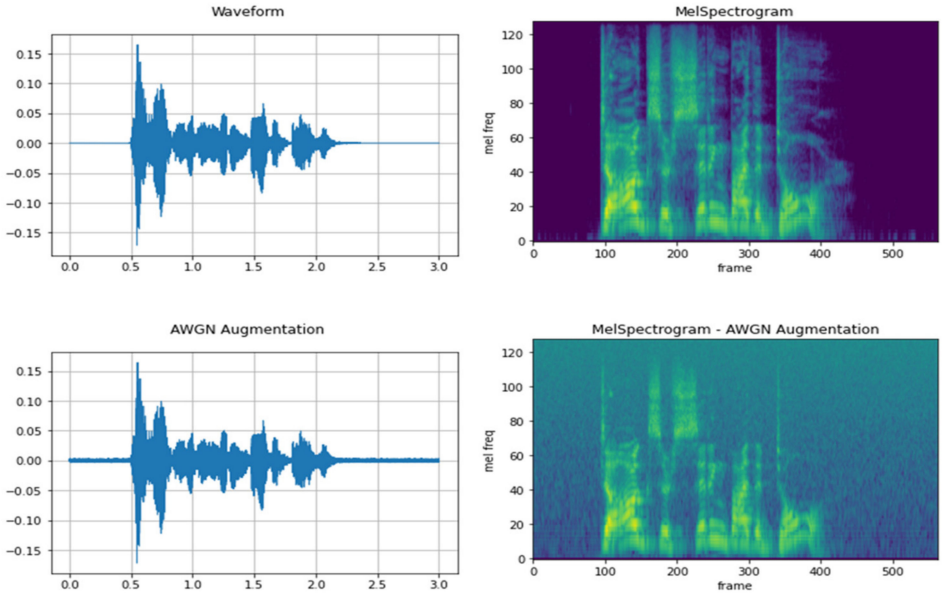

Figure 2. Comparison of the original audio signal's waveform and mel-spectrogram compared with an Additive White Gaussian Noise augmented signal.

3.2.2. SpecAugment

SpecAugment applies time masking and frequency masking on a log-mel spectrogram. This reduces overfitting during training and improves the model's generalization, as models trained with SpecAugment become more invariant to small variations in acoustic features [32]. The following masking is implemented as follows:

1. Frequency masking: a parameter value F where a masking size f belongs to a uniform distribution from 0 to F is selected. These consecutive frequencies are masked with values of 0
2. Time masking: a parameter value T where the masking size t belongs to a uniform distribution from 0 to T is selected. These consecutive time steps are masked with values of 0.

Although considered to be a simple technique, adaptation of SpecAugment has been reported to provide relative improvements in the domains of speech recognition [33], speaker verification systems [34], and speech emotion recognition [35]. However, it there were different experiments done using SpecAugment, they would show that time masking has minimum-to-no benefits on speech related classification tasks, while frequency masking provides better results in augmenting the training data for such tasks. A comparison of the original audio signal's waveform and mel-spectrogram compared with the SpecAugment augmented signal is shown on Figure 3.

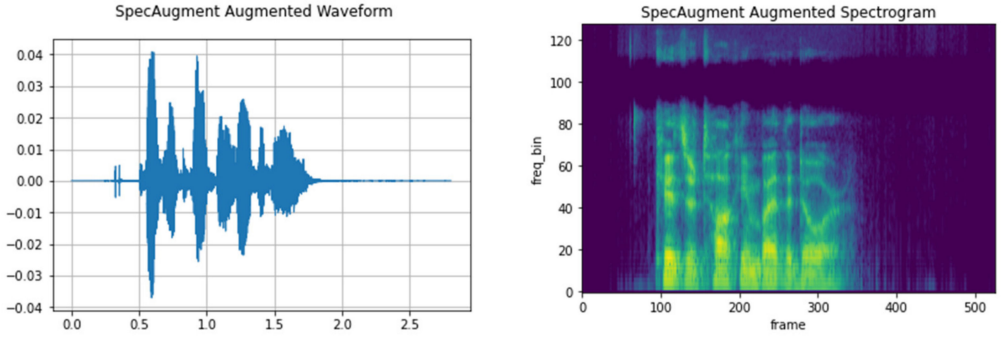

Figure 3. Waveform and Spectrogram representations of an SpecAugment augmented sample.

3.2.3. Room Impulse Response

Room Impulse Response (RIR) is the transfer function between the sound source and microphone. An impulse response of a dynamic system describes how it reacts when presented with a brief input signal called response. The reaction of the system can be influenced by the room's surroundings. The impulse signal contains the frequencies which capture a microphone's position and reverberations. Samples recorded on a studio such as that of RAVDESS dataset can be convoluted with an impulse response to simulate the audio files as if it was recorded in different scenarios. In this paper, the dataset was randomly convoluted with random IR samples from the EchoThief Impulse Response Library [36]. A comparison of the original audio signal's waveform and mel-spectrogram compared with the RIR augmented signal is shown on Figure 4.

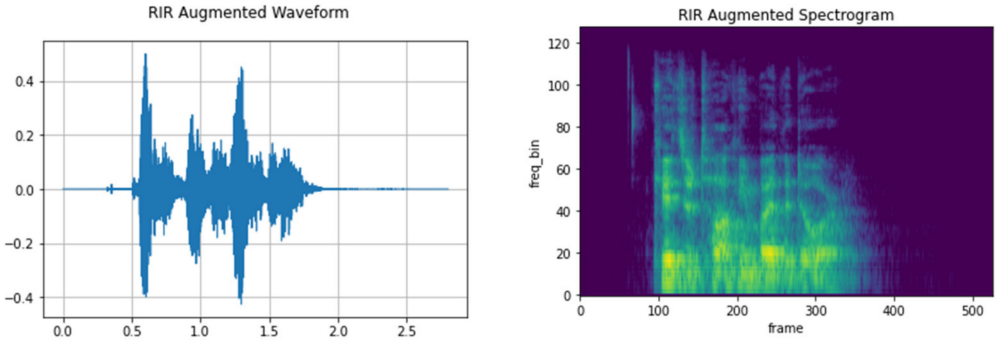

Figure 4. Waveform and Spectrogram representations of a Room Impulse Response augmented sample.

3.2.4. Tanh Distortion

Tanh distortion technique involves using a mathematical function that directly modifies the values of the audio signal. This provides a rounded soft clipping kind of distortion amount that is proportional to the loudness of the input and the pre-gain. Since the tanh function is symmetric, the positive and negative parts of the signal are squashed the same way. This distortion technique generally adds harmonics to the signal, changing the timbre of the sound. A comparison of the original audio signal's waveform and mel-spectrogram compared with the tan distorted signal is shown on Figure 5.

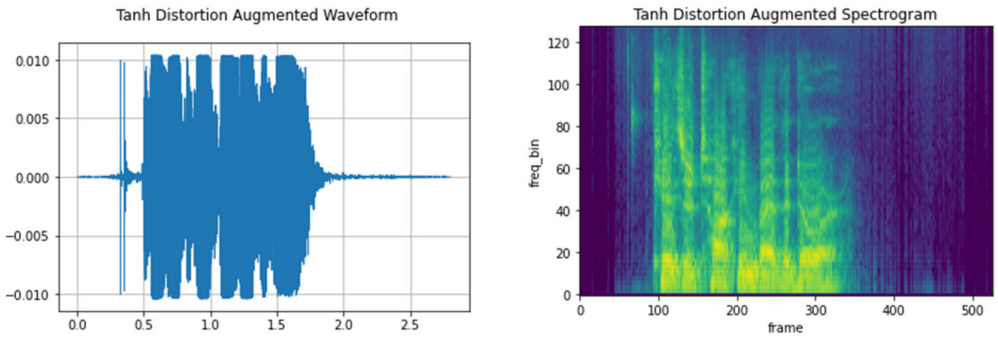

Figure 5. Waveform and Spectrogram representations of a Tan Distorted sample.

4. Experiments

4.1. Pre Processing Details

First, the speech utterances from the RAVDESS dataset were used in the experiments. A total of 1440 speech data are used prior to augmentation. The dataset is divided into training, validation, and testing subsets with the ratio of 8:1:1. All of the speech data were transformed into mel spectrograms with a sample rate of 48,000. Mel spectrogram is produced from STFT frames with a mean on each column resulting in a matrix that produces a feature array. The python library Librosa [37] was used to extract mel spectrogram features

with an FFT window length of 1024, hamming window of length 512, hop size of 256, and 128 mel bins as its parameters.

The training dataset is then augmented using different augmentation techniques for multi folds of N. In this experiment, and N sizes of 1 and 2 are used to augment the training signals with the same augmentation technique by a single (increases training data twice) or a two-fold (increases training data thrice) augmentation.

On the AGWN augmentation, a minimum sound-to-noise (SNR) ratio of 15, and maximum of 30 before performing a mel spectrogram transform. During the SpecAugment augmentation, the signal is transformed with the masking parameter of 40 on the frequency parameter f after a mel spectrogram transformation. Signals augmented with an RIR are convoluted with random IR signals from the EchoThief Impulse Response Library. Finally, a Tanh Distortion with a randomized amount of distortion from a uniform distribution between 0.01 and 0.5 is applied on the signal. Augmentations were performed using Audiomentations.

4.2. Model Implementation

Two main models (Parallel CNN-BLSTM-Attention, and Parallel CNN-Transformer models) used in this paper were both inspired by the 2D convolutional blocks of the LeNet [38] architecture and implemented using PyTorch. The LeNet's 2D convolutional blocks were selected because of their plain architecture that could be further improved by implementing additional techniques such as cross connections, inceptions modules, or residuals connections.

In the Parallel CNN-BLSTM-Attention model, a bidirectional LSTM layer is designed with an attention layer, paralleled to a 4-layer deep 2D convolutional block. The convolution layers take an input with the format of batch size, channel, height, and width. The mel spectrogram input feature has a shape of (N, 1, 128, 563), where the N is the number of training data. A single channel is used mainly because a mel spectrogram feature can be considered a black-and-white image instead of the usual 3-channeled RGB image. The channel consists of the intensities for the 128 mel frequency bands at 563 timesteps. The training of this model has an epoch of 500.

In the Parallel CNN- Transformer model, a transformer encoder is designed parallel with it as a 3-layer deep 2D convolutional block. The input feature also takes the shape of (N, 1, 128, 563) similar with the Paralleled CNN-BLSTM- Attention model. The first layer takes a $1 \times 3 \times 3$ filter producing an output of 16 channels, with a batch normalization applied to the output feature map before using an ReLU activation. A max-pooling layer is applied after the activation, followed by a drop-out layer with the probability of 0.3 on all subsequent layers. The second layer expands the output feature map to a depth of 32 channels, while increasing the max pool kernel size. Finally, the third convolutional block bottlenecks the output back into a feature map volume of 16. Two 3-layered CNN blocks are implemented parallel with a Transformer Encoder layer. The Transformer-Encoder layer is inspired by [39], with the goal to help predict frequency distributions of the emotion based on the global structure of the mel spectrogram per emotion. Using the multi-head self-attention layer of the transformer, the network takes into consideration multiple previous time steps when predicting the next. The training on this model has an epoch of 100. The overall specifications of the Parallel CNN-Transformer and Parallel CNN-BLSTM+Attention network are shown in Tables 3 and 4, respectively.

The parallel model trained on non-augmented audio from the RAVDESS dataset is compared with CNN based models such as VGG-16 and ResNet50, as well as with standalone LSTM-Attention and Transformer models. We have also experimented on a network of time-distributed CNN layers stacked on a transformer.

Both the proposed parallel models are compared with a standard VGG-16 model as a baseline. These parallel models effectively improve performance without additional auxiliary networks that use complex architectures, while keeping a relatively smaller

number of training parameters compared with traditional CNN based networks and time-distributed CNN ensemble networks.

Table 3. Overall specification of Parallel-CNN-Transformer network with set of layer sizes, output sizes, and number of units.

Layer	Output Channels	No. of Units
Conv2d_1_1	16	[3 × 3, stride 1, padding 1] [batch normalization] [2 × 2 max pooling, stride 2] [dropout 0.3]
Conv2d_1_2	32	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
Conv2d_1_2	64	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
Conv2d_2_1	16	[3 × 3, stride 1, padding 1] [batch normalization] [2 × 2 max pooling, stride 2] [dropout 0.3]
Conv2d_2_2	32	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
Conv2d_2_2	64	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
Transformer Encoder	192	[encoder_layers, attention_heads 4, FC 512] × 4 layers [dropout 0.4]
Output	8	[2 * 128 + 256 FC layers] softmax

Table 4. Overall specification of Parallel-CNN-BLSTM-Attention network with set of layer sizes, output sizes, and number of units.

Layer	Output Channels	No. of Units
Conv2d_1	16	[3 × 3, stride 1, padding 1] [batch normalization] [2 × 2 max pooling, stride 2] [dropout 0.3]
Conv2d_2	32	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
Conv2d_3	64	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
Conv2d_4	64	[3 × 3, stride 1, padding 1] [batch normalization] [4 × 4 max pooling, stride 4] [dropout 0.3]
BLSTM Block	256	[2 × 4 maxpooling, stride 2 × 4] [128 bi-directional lstm] [dropout 0.1] [2 * 128, attention layer]
Output	8	[256 + 256 FC layers] softmax

5. Results and Discussion

The experiments were performed on a desktop computer with the following configurations: Intel Core i7-10900K@3.70 Ghz, 64 GB RAM, and NVIDIA Quadro M6000 with 24 GB RAM. The Baseline model of VGG-16 was trained with an epoch of 100, having an accuracy score of 61.54 without any augmentation techniques applied. Applying an Additive White Gaussian Noise technique achieved the highest accuracy score of 76.92 on the said model with a single-fold augmentation. This is followed by the Tanh Distortion at 71.68, SpecAugment at 67.13, and Room Impulse Response at 62.94. For the two-fold augmentation experiment, the AWGN technique also achieved the highest accuracy score of 80.77, followed by Tanh Distortion at 78.32, SpecAugment at 71.33, and RIR at 68.53. With these as our baseline accuracy scores, we compared them with the proposed parallel models: Parallel LSTM-Attention-CNN model, and Parallel Transformer-CNN network trained on the RAVDESS dataset.

On the Parallel LSTM-Attention-CNN model, experiments have shown that applying no augmentation on the training data achieved an accuracy of 65.94, which is at least 4.4% higher than the baseline. However, application of SpecAugment on the training data achieved the highest accuracy score on a single fold augmentation experiment with a score of 81.33, in contrast of AWGN on the baseline model. These are followed by the Tanh Distortion at 74.33, RIR at 71.33, and AWGN at 69.00. On the two-fold augmentation experiment, SpecAugment also achieved the highest accuracy score of 85.67, followed by AWGN, Tanh Distortion, and RIR at 85.27, 84.89, and 70.22, respectively. It is quite interesting to see that the frequency based SpecAugment technique has significantly increased the accuracy of the model on both single and two-fold augmentation experiments.

On the Parallel Transformer-CNN model, an accuracy of 81.33 is achieved without using any data augmentation techniques. This is significantly higher than the accuracy scores from the baseline and the LSTM-Attention-CNN networks. During the single fold augmentation experiment the model achieved an accuracy of 84.80 using an AWGN data augmentation technique, followed by SpecAugment and Tanh Distortion, both at 82.85, which are also quite similar with RIR augmentation at 82.80. On the two-fold augmentation experiment, the highest accuracy score of 89.33 was attained using a Tanh Distortion augmentation; these are followed by AWGN with 88.89, SpecAugment with 84.86, and RIR at 83.22

Comparing the parallel models on training data without additional augmentation, smaller standalone LSTM+Attention and Transformer networks of the similar architecture of our parallel network achieved an accuracy of 60.00 and 70.67, respectively. These are much more efficient compared with traditional CNN networks, as they have significantly smaller numbers of training parameters compared with as VGG-16 and ResNet50 which achieved an accuracy of 61.54 and 69.33, respectively.

Networks with a time-distributed CNN, stacked with BLSTM+Attention and Transformer layers instead of running on parallel, were also experimented, with accuracies of 64.67 and 77.33. Although these networks have accuracies similar to our proposed network, they suffer from a large number of training parameters making them more computationally intensive when deployed on systems. A summary of experiments of all the model architectures' metrics and parameter sizes is shown in Table 5.

Table 5. Weighted average precision, sensitivity, F1 score, and accuracy of different models compared with parallel networks.

Model Architecture	Trainable Parameters	Metrics			
		Precision	Sensitivity	F1	Accuracy
VGG-16 (baseline)	57,092,000	0.58	0.62	0.59	61.54
ResNet50	23,524,424	0.70	0.69	0.69	69.33
LSTM + Attention	200,969	0.62	0.60	0.60	60.00
Transformer	268,816	0.71	0.71	0.71	70.67
Time-distributed CNN + BLSTM + Attention	582,761	0.65	0.65	0.65	64.67
Time-Distributed CNN + Transformer	10,312,984	0.78	0.77	0.77	77.33
Parallel CNN + BLSTM + Attention (ours)	261,288	0.62	0.71	0.71	65.94
Parallel CNN + Transformer (ours)	395,176	0.80	0.82	0.80	81.33

The summary of all accuracy scores is shown in Table 6, while the confusion matrix for the proposed models is shown on Figure 6.

Table 6. Weighted average precision, sensitivity, F1 score, and accuracy of each model augmented with different augmentation techniques on test data.

Augmentation Technique		VGG-16 (Baseline)				Parallel BLSTM-Attention-CNN				Parallel Transformer CNN			
		Precision	Sensitivity	F1	Acc	Precision	Sensitivity	F1	Acc	Precision	Sensitivity	F1	Acc
No Augmentation		0.58	0.62	0.59	61.54	0.62	0.71	0.59	65.94	0.80	0.82	0.80	81.33
Single Fold Augmentation	AWGN	0.72	0.78	0.71	76.92	0.67	0.81	0.67	69	0.82	0.84	0.81	84.80
	RIR	0.60	0.63	0.61	62.94	0.72	0.73	0.71	71.33	0.75	0.78	0.75	82.80
	SpecAugment	0.62	0.68	0.63	67.13	0.82	0.86	0.82	81.33	0.76	0.8	0.76	82.85
	Tanh Distortion	0.72	0.82	0.71	71.68	0.74	0.82	0.74	74.33	0.68	0.77	0.68	82.85
Two-Fold Augmentation	AWGN	0.78	0.82	0.78	80.77	0.84	0.88	0.83	85.27	0.88	0.89	0.88	88.89
	RIR	0.69	0.71	0.69	68.53	0.69	0.80	0.71	70.22	0.73	0.78	0.71	83.22
	SpecAugment	0.72	0.74	0.71	71.33	0.87	0.86	0.86	85.67	0.8	0.86	0.81	84.86
	Tanh Distortion	0.74	0.82	0.75	78.32	0.69	0.81	0.71	84.89	0.89	0.91	0.96	89.33

It can also be noted through the confusion matrices on Figure 6 that generally the models often get confused with classifying the emotion category sad, by misclassifying it as either neutral or calm, as it had the most number of false positives which is evident on the Parallel CNN-Transformer networks with two-fold AWGN and SpecAugment, as well as on the Parallel CNN-BLSTM-Attention networks with the RIR augmentation on both the single and two-fold augmentation experiments. The models also have a little bit of difficulty identifying the category calm with neutral, which can be observed from the confusion matrices. This might be due to the two categories being closely related with each other.

Comparing all the accuracy scores, simple traditional augmentation techniques that are not computationally expensive seemed to provide more improvement to the model, as compared with a more computationally expensive approach such as RIR augmentation. Although RIR augmentation has provided improved accuracy scores on all models, due to the complexity of different IR samples on the dataset doubling the augmentations with such technique does not provide any significant increase in the performance.

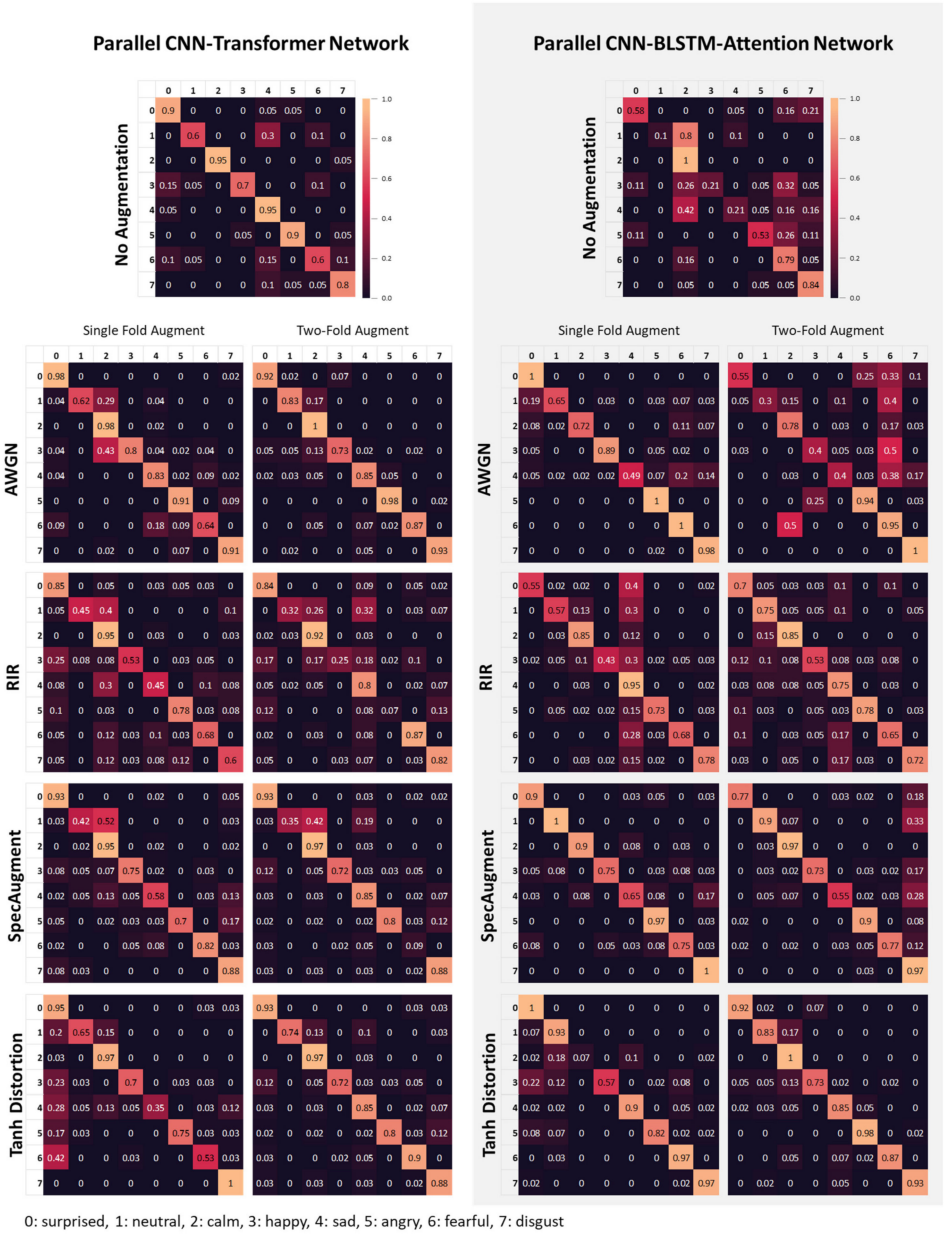

Figure 6. Normalized Confusion Matrix for the Parallel CNN-Based Networks.

6. Conclusions

Overall, using CNN networks parallel with time-sequence models such as Transformers and LSTM with attention has shown good improvement on an SER task using the RAVDESS dataset. As time and sequence are quite important for tasks involving speech, these techniques provide better feature representation of audio data compared to deep stacked layers of models such as VGG-16 and ResNet, as well as with standalone LSTM+Attention and Transformer networks. Further, experiments showed that training CNN layers along with LSTM+Attention and Transformers in parallel is more efficient than stacked CNN layers with a time-distributed wrapper, as they performed better with accuracies higher than stacked networks while using a significantly smaller number of trainable parameters. The experiments on multi-fold data augmentations also showed great improvement on the SER classification task by increasing the training data, making them less prone to overfitting and thus making it more robust. For the types of data augmentation, simple augmentation techniques such as AWGN and Tanh Distortion could provide a simple, yet quite significant, increase in the performance of the models, rather

than that of RIR based augmentation techniques which cost more in computational power to augment data but do not significantly provide a better performance compared with models trained without applying augmentation.

Future research can be directed on larger N fold augmentations, more complex spectral based augmentations, and much more advanced augmentation techniques using deep learning approach such as Autoencoders, and Generative Adversarial Networks (GANS). The parallel models can be further developed by experimenting on additional residual skip connections, or adding auxiliary networks such as Wav2Vec and feature mapping networks for a completely end-to-end network architecture. Further research is also expected on different emotional speech datasets and on mixed augmentation techniques.

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

Funding: This research and APC was funded by the Industrial Technology Innovation Program (No. 20012603, Development of Emotional Cognitive and Sympathetic AI Service Technology for Remote (Non-face-to-face) Learning and Industrial Sites) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

Data Availability Statement: The data presented in this study are openly available in https://zenodo.org/record/1188976 at https://doi.org/10.1371/journal.pone.0196391.

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

References

1. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
2. Park, S.-M.; Kim, Y.-G. A Metaverse: Taxonomy, Components, Applications, and Open Challenges. *IEEE Access* **2022**, *10*, 4209–4251. [CrossRef]
3. Chen, M.; Zhou, P.; Fortino, G. Emotion Communication System. *IEEE Access* **2016**, *5*, 326–337. [CrossRef]
4. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* **2019**, *7*, 117327–117345. [CrossRef]
5. 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]
6. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
7. Mao, X.; Chen, L.; Fu, L. Multi-level speech emotion recognition based on HMM and ANN. In Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 31 March–2 April 2009; Volume 7, pp. 225–229. [CrossRef]
8. Cheng, X.; Duan, Q. Speech Emotion Recognition Using Gaussian Mixture Model. In Proceedings of the 2012 International Conference on Computer Application and System Modeling (ICCASM 2012), Taiyuan, China, 27–29 July 2012; pp. 1222–1225. [CrossRef]
9. Lanjewar, R.B.; Mathurkar, S.; Patel, N. Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) Techniques. *Procedia Comput. Sci.* **2015**, *49*, 50–57. [CrossRef]
10. Jain, M.; Narayan, S.; Balaji, K.P.; Bharath, K.; Bhowmick, A.; Karthik, R.; Muthu, R.K. Speech Emotion Recognition using Support Vector Machine. *arXiv* **2020**, arXiv:2002.07590. [CrossRef]
11. Al Dujaili, M.J.; Ebrahimi-Moghadam, A.; Fatlawi, A. Speech emotion recognition based on SVM and KNN classifications fusion. *Int. J. Electr. Comput. Eng. (IJECE)* **2021**, *11*, 1259–1264. [CrossRef]
12. Harár, P.; Burget, R.; Dutta, M.K. Speech emotion recognition with deep learning. In Proceedings of the 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Delhi, India, 2–3 February 2017; pp. 137–140. [CrossRef]
13. Fahad, S.; Deepak, A.; Pradhan, G.; Yadav, J. DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features. *Circuits Syst. Signal Process.* **2020**, *40*, 466–489. [CrossRef]
14. Poon-Feng, K.; Huang, D.Y.; Dong, M.; Li, H. Acoustic emotion recognition based on fusion of multiple feature-dependent deep Boltzmann machines. In Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, Singapore, 12–14 September 2014; pp. 584–588. [CrossRef]

15. Qayyum, A.B.A.; Arefeen, A.; Shahnaz, C. Convolutional neural network (CNN) based speech-emotion recognition. In Proceedings of the 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, 28–30 November 2019; pp. 122–125. [CrossRef]
16. 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. [CrossRef]
17. Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B. Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1675–1685. [CrossRef]
18. Atmaja, B.T.; Akagi, M. Speech emotion recognition based on speech segment using LSTM with attention model. In Proceedings of the 2019 IEEE International Conference on Signals and Systems, Kuala Lumpur, Malaysia, 18–19 September 2019; pp. 40–44. [CrossRef]
19. Zeng, Y.; Mao, H.; Peng, D.; Yi, Z. Spectrogram based multi-task audio classification. *Multimed. Tools Appl.* **2017**, *78*, 3705–3722. [CrossRef]
20. Popova, A.S.; Rassadin, A.G.; Ponomarenko, A.A. Emotion recognition in sound. In Proceedings of the International Conference on Neuroinformatics, Moscow, Russia, 2–6 October 2017; Springer: Cham, Switzerland, 2017; pp. 117–124. [CrossRef]
21. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]
22. Li, H.; Ding, W.; Wu, Z.; Liu, Z. Learning fine-grained cross modality excitement for speech emotion recognition. *arXiv* **2020**, arXiv:2010.12733. [CrossRef]
23. Lu, Z.; Cao, L.; Zhang, Y.; Chiu, C.-C.; Fan, J. Speech Sentiment Analysis via Pre-Trained Features from End-to-End ASR Models. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020. [CrossRef]
24. Pepino, L.; Riera, P.; Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv* **2021**, arXiv:2104.03502. [CrossRef]
25. Cai, X.; Yuan, J.; Zheng, R.; Huang, L.; Church, K. Speech Emotion Recognition with Multi-Task Learning. *Interspeech* **2021**, *2021*, 4508–4512. [CrossRef]
26. Han, S.; Leng, F.; Jin, Z. Speech emotion recognition with a ResNet-CNN-Transformer parallel neural network. In Proceedings of the 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 14–16 May 2021; pp. 803–807. [CrossRef]
27. Slimi, A.; Nicolas, H.; Zrigui, M. Hybrid Time Distributed CNN-Transformer for Speech Emotion Recognition. In Proceedings of the 17th International Conference on Software Technologies ICSOFT, Lisbon, Portugal, 11–13 July 2022. [CrossRef]
28. Xia, Y.; Chen, L.-W.; Rudnicky, A.; Stern, R.M. Temporal Context in Speech Emotion Recognition. *Interspeech* **2021**, *2021*, 3370–3374. [CrossRef]
29. Wei, S.; Zou, S.; Liao, F.; Lang, W. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *J. Phys. Conf. Ser.* **2020**, *1453*, 12085. [CrossRef]
30. Praseetha, V.M.; Joby, P.P. Speech emotion recognition using data augmentation. *Int. J. Speech Technol.* **2021**, 1–10. [CrossRef]
31. Huang, C.; Chen, G.; Yu, H.; Bao, Y.; Zhao, L. Speech emotion recognition under white noise. *Arch. Acoust.* **2013**, *38*, 457–463. [CrossRef]
32. 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. *arXiv* **2019**, arXiv:1904.08779. [CrossRef]
33. Park, D.S.; Zhang, Y.; Chiu, C.C.; Chen, Y.; Li, B.; Chan, W.; Wu, Y. Specaugment on large scale datasets. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6879–6883. [CrossRef]
34. Faisal, M.Y.; Suyanto, S. SpecAugment impact on automatic speaker verification system. In Proceedings of the 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 5–6 December 2019; pp. 305–308. [CrossRef]
35. Cui, X.; Goel, V.; Kingsbury, B. Data Augmentation for Deep Neural Network Acoustic Modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1469–1477. [CrossRef]
36. Warren, C. Echothief Impulse Response Library. Available online: http://www.echothief.com/ (accessed on 31 August 2022).
37. 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. [CrossRef]
38. LeCun, Y. LeNet-5, Convolutional Neural Networks. 2015. Available online: http://yann.lecun.com/exdb/lenet (accessed on 31 August 2022).
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]