sensors

MDPI

Article

Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning

Apeksha Aggarwal [1,†], Akshat Srivastava [2,†], Ajay Agarwal [3], Nidhi Chahal [4], Dilbag Singh [5], Abeer Ali Alnuaim [6], Aseel Alhadlaq [6] and Heung-No Lee [5,*]

1 Department of Computer Science Engineering & Information Technology, Jaypee Institute of Information Technology, A 10, Sector 62, Noida 201307, India; apeksha.aggarwal@mail.jiit.ac.in or apeksha.aggarwal785@gmail.com
2 School of Computer Science Engineering and Technology, Bennett University, Plot Nos 8-11, TechZone 2, Greater Noida 201310, India; srivastavaakshat8@gmail.com
3 Department of Information Technology, KIET Group of Institutions, Delhi-NCR, Meerut Road (NH-58), Ghaziabad 201206, India; ajay.agarwal@kiet.edu
4 Nidhi Chahal, NIIT Limited, Gurugram 110019, India; nidhi.vce@gmail.com
5 School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea; dilbagsingh@gist.ac.kr or dggill2@gmail.com
6 Department of Computer Science and Engineering, College of Applied Studies and Community Services, King Saud University, P.O. Box 22459, Riyadh 11495, Saudi Arabia; abalnuaim@ksu.edu.sa (A.A.A.); asalhadlaq@ksu.edu.sa (A.A.)
* Correspondence: heungno@gist.ac.kr
† These authors contributed equally to this work.

check for updates

Citation: Aggarwal, A.; Srivastava, A.; Agarwal, A.; Chahal, N.; Singh, D.; Alnuaim, A.A.; Alhadlaq, A.; Lee, H.-N. Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors* **2022**, *22*, 2378. https://doi.org/10.3390/s22062378

Academic Editors: Piyush Kumar Shukla, Ahmed A. Abd El-Latif, Rupak Kharel, Udai Pratap Rao, Prashant Kumar Shukla, Raffaele Gravina and Raffaele Bruno

Received: 31 January 2022
Accepted: 15 March 2022
Published: 19 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: Recognizing human emotions by machines is a complex task. Deep learning models attempt to automate this process by rendering machines to exhibit learning capabilities. However, identifying human emotions from speech with good performance is still challenging. With the advent of deep learning algorithms, this problem has been addressed recently. However, most research work in the past focused on feature extraction as only one method for training. In this research, we have explored two different methods of extracting features to address effective speech emotion recognition. Initially, two-way feature extraction is proposed by utilizing super convergence to extract two sets of potential features from the speech data. For the first set of features, principal component analysis (PCA) is applied to obtain the first feature set. Thereafter, a deep neural network (DNN) with dense and dropout layers is implemented. In the second approach, mel-spectrogram images are extracted from audio files, and the 2D images are given as input to the pre-trained VGG-16 model. Extensive experiments and an in-depth comparative analysis over both the feature extraction methods with multiple algorithms and over two datasets are performed in this work. The RAVDESS dataset provided significantly better accuracy than using numeric features on a DNN.

Keywords: speech emotion recognition; machine learning; neural network

1. Introduction

Speech is the most fundamental method of human communication. Humans can naturally detect the emotion in the speech they are presented with. However, it is not so straightforward for machines. This is where the importance of speech emotion recognition (SER) is highlighted. An SER system utilizes files containing speech data and classifies them into various emotions irrespective of the semantic contents [1]. A variety of emotions are generally classified by SER, including "happy", "sad", "angry" and "neutral". SER systems aim to give rise to efficient methods of detecting emotions; this function is crucial for adding human elements like emotional response in machines. A well-working SER system can be utilized in several fields involving human-machine communication, ranging from mobile phone use to driving cars and beyond. Technologies like SER systems are

getting increasingly important to reduce human time and effort. SER systems employ machines or robots to have a meaningful dialogue between a human and a machine. To create such a system, machine learning and deep learning models can be employed. This involves extracting significant features from raw data and utilizing them to make machines understand human emotions via model training [2]. During the training process, the model learns to classify information and produce desired outputs while maintaining a certain level of accuracy. Given the sheer number of options we can work with while crafting such a model, we have a huge number of permutations to try.

The task of emotion recognition from speech is divided into two major sections: feature selection and extraction, and classification. The authors in [3] have utilized a total of five features: Mel-Frequency Cepstral Coefficients (MFCC), Mel, Chroma, Tonnetz, and various permutations were tried and tested. The authors in [4] made use of a Decision Tree Classifier and a Convolutional Neural Network (CNN) to approach this problem. Authors extracted melcepst coefficients, randomize the data and then proceeded to train and test them over both the Decision Tree Classifier and the CNN. The highest accuracy that was achieved in this work was 72%. Similarly, the authors in [5] used the Random Forest Classifier and the Decision Tree Classifier to tackle the problem at hand. The best average recognition rate they reached was 66.28% by utilizing decision trees. Recently, authors in [6] also used Mel-Frequency Cepstrum Coefficients for recognizing emotions using speech, which have proven to be a crucial feature set for audio data.

Another effective and popular way to extract emotions from speech was to use CNNs [7,8] and Deep CNNs [9,10]. The authors in [11] used the Deep CNNs to approach the task. They managed the best test accuracy of 40.2% using two convolutional layers and two pooling layers to train and test the data. Support Vector Machines remained a popular choice in classification problems and the authors in [12] had done a comprehensive study utilizing the same. They used a biased SVM to approach the problem at hand. They achieved a maximum average accuracy of 58.24%.

Apart from CNNs [13–15], Long Short-term Memory Networks (LSTM) [16], DNN [17,18] and CNN-LSTM hybrids have also shown promising results in the field of emotion recognition. LSTMs are more advanced Recurrent Neural Networks (RNN) optimized to use gates to control information flow. Pandey et al. [19] adopted this methodology for the experiments. They extracted magnitude spectrograms, log-mel spectrograms, and MFCC and used a CNN-LSTM hybrid to train and test the data. The authors achieved an accuracy of 82.35% by using MFCCs as their input. However, only four emotions were tested in this work.

The authors in [20] utilized deep neural networks to calculate the weighted and unweighted accuracy of the model. This approach yielded a maximum weighted accuracy of 70.1% and maximum unweighted accuracy of 60.7%. Attention pooling is another option to tackle SER. The authors in [21] utilized attention pooling-based CNN and managed a weighted accuracy of 71.75% and an unweighted accuracy of 68.06%. Palo et al. [22] proposed another technique in which authors extracted different features and calculated the maximum accuracy achieved. The authors in this work achieved the best performance by using MFCCs when only four emotions (bored, angry, sad, and surprised) were taken into account. A similar study was done in [23] where authors utilized Log Mel-Spectrograms, MFCCs, eGeMAPS, and Prosody. Log Mel-Spectrograms gave the best accuracy when using a CNN (55.92%) and MFCCs showed the best accuracy when using an Attentive CNN (56.10%). In another work [24], different ANN models were used to calculate the accuracy of emotion prediction. The authors achieved the best results by using a frame-based CNN network, which gave a test accuracy of 64.78%. Lech et al. [1] proposed SER system which utilized AlexNet; a pre-trained image classification network and achieved the average accuracy of approximately 80%.

Another popular approach to the problem is making use of RNNs of various types, one such approach was proposed in [25], where authors utilized a CNN-RNN hybrid (CRNN) to address speech emotion recognition. They achieved a maximum unweighted accuracy

of 63.98% by utilizing an HSF-CRNN system. In another experiment involving an RNN, the authors at [26] used an LSTM Neural Network to achieve an unweighted accuracy of 60.02%. The authors in [27–29] used similar RNN and LSTM systems to approach SER, achieving accuracies of 63.5%, 58.7%, and 63.89%, respectively.

For the present work, we have worked upon one-dimensional and two-dimensional data. For one dimensional data, we have used three different classifiers: Decision Tree [30], Random Forest [31], and MLP [32]. For two-dimensional data, although the primary focus has been on deep neural networks, we have also proposed the use of a dummy classifier to set a baseline accuracy. We have generated mel spectrograms [33] as 2D features for this research in addition to the generation of 1-D features from the audio dataset to test the model on four specific emotions: happy, sad, angry, and neutral. The contribution of this paper are summarized as follows:

1. Two-way feature extraction is proposed by utilizing super convergence to extract two sets of potential features from the speech data.
2. Principal component analysis (PCA) and deep neural network (DNN) with dense and dropout layers are applied to the features obtained from the proposed two-way feature extraction model.
3. The pre-trained VGG-16 model is also trained on the features from the proposed two-way feature extraction model.
4. Multimodal speech data is utilized for training.

2. Materials and Methods

In this work, we have proposed a two-way approach to extract features from the speech dataset. Schematic representations of the steps are depicted in Figure 1. Both the approaches are described further in this section.

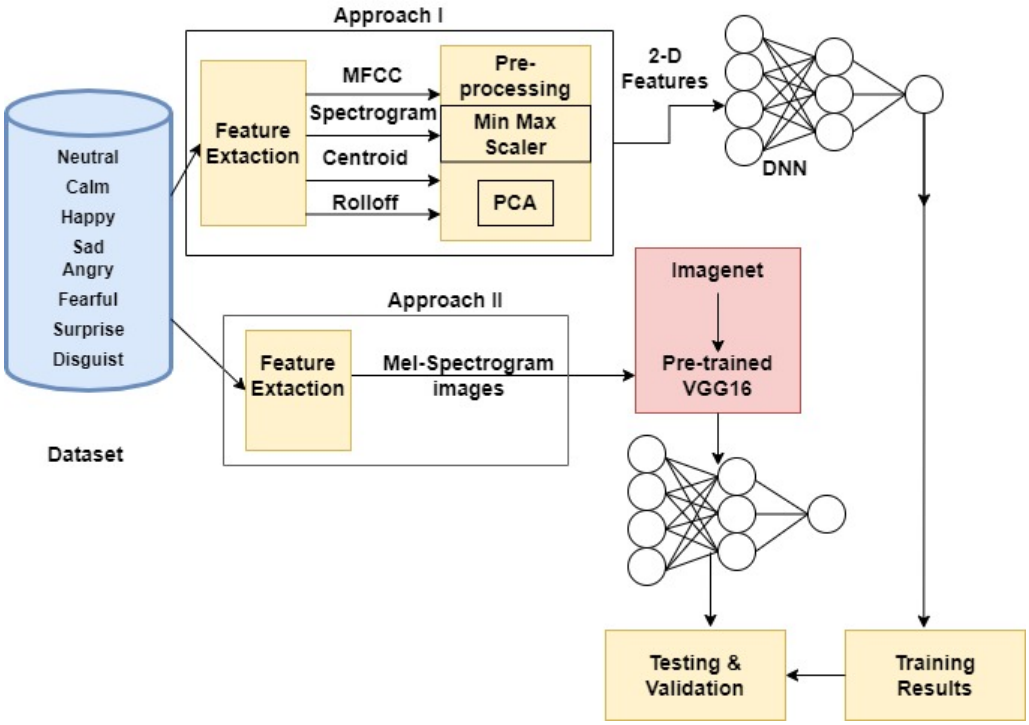


Figure 1. Schematic representation of two-way feature extraction for speech data.

2.1. Dataset Description

For this research work, two datasets have been utilized. First is the Toronto Emotional Speech Set (TESS) [34] and the second one is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [35]. TESS consists of a set of 200 target words; every target word is spoken post the phrase “Say the word". The same has been recorded by two actresses portraying seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are a total of 2800 data samples in the dataset. RAVDESS consist of 24 actors, 60 trials per actor. A total of eight emotions are portrayed (calm, happy, sad, angry, fearful, surprise, and disgust) in this dataset. There are a total of 1440 data samples in the dataset.

2.2. Approach I

In the present work we have proposed two approaches to extract two types of features. In the first approach we work directly on the audio dataset to obtain numerical features. This section describes the details further.

2.2.1. Feature Extraction

To anticipate the emotion of a given speech, we need to identify and extract one or more meaningful features. An audio dataset is rendered suitable by extracting suitable features from voice signals. In this approach, a combination of MFCC, Log Mel-Spectrogram, Chroma, Spectral centroid, and Spectral rolloff have been extracted. The librosa library has been used for feature extraction from audio signals directly. Post the extraction, the features of each file along with the labels have been converted to a 2-D feature vector.

2.2.2. Dimensionality Reduction and Preprocessing

A total of 180 features have been extracted from the audio files. To remove the sparsity and high dimensionality of the dataset, further pre-processing of the data have been performed. Firstly, the data have been normalized using the MinMaxScaler from sklearn. Furthermore, to reduce the dimensionality, PCA has been utilized. PCA is used as it reduced the overfitting significantly by eliminating highly correlated variables. Using PCA, 80 important features have been selected to allow for effective training and testing.

2.2.3. Model Architecture

In the first approach, DNN [36,37] is used along with dropout layers. The architecture for the DNN is shown in Figure 2. We have used a 13 layers for the network architecture viz. Seven dense layers and six dropout layers. For the dense layers, 1024 units have been used in the input layer and 512, 256, 128, 64, and 32 units have been used for the rest of the hidden layers, respectively. Since there are eight classes taken into consideration, the output layer consists of eight units. For the input and hidden layers, ‘relu’ has been used as the activation function, and for the output layer, the ‘softmax’ activation function [37,38] is utilized. Each of the input and hidden layers is followed by a dropout layer with a rate of 0.2. The model has been trained for a total of 400 epochs, with ‘rmsprop’ as the optimizer and ‘categorical crossentropy’ as the loss function.

2.3. Approach II

In this section, the second approach for feature extraction is described. In this approach, we have utilized spectrograms as image features.

2.3.1. Feature Extraction

In this approach, we have used another method of feature extraction by employing transfer learning. In this approach, we have generated Log Mel-Spectrogram images from the input audio dataset. Using the librosa library, the Mel-Spectrogram image for each file has been extracted and saved, respectively, to the particular emotion class.

In summary, a total of 1440 and 2800 images have been extracted for RAVDESS and TESS datasets, respectively.

2.3.2. Model Architecture

We have further utilized VGG16's *fastai* implementation. VGG16 is a Convolution Neural Network-based transfer learning model. It makes use of Conv2d layers along with BatchNorm2d and MaxPool2d layers. There are a total of 16 layers in VGG16.

2.4. Experiments

For the first approach, the data have been split into training and testing data with a train_size of 0.8 and test_size of 0.2. The extracted numeric features have been passed on to this DNN for training. Dropout layers have been used to prevent any overfitting. Furthermore, the target labels have been label encoded into categorical values. For constructing the model, we have utilized the Keras library. After every Dense layer, a dropout layer has been added. Softmax has been used as the activation function and categorical cross-entropy has been used as the loss function. The data have been trained for a total of 25 epochs for the TESS dataset and 400 epochs for the RAVDESS dataset, respectively.

For the second approach, the model was loaded with imagenet weights and the data have been trained for a total of five epochs for the TESS dataset and 25 epochs for the RAVDESS dataset.

Layer (type)	Output Shape	Param #
dense_15 (Dense)	(None, 1024)	82,944
dropout_6 (Dropout)	(None, 1024)	0
dense_16 (Dense)	(None, 512)	524,800
dropout_7 (Dropout)	(None, 512)	0
dense_17 (Dense)	(None, 256)	131,328
dropout_8 (Dropout)	(None, 256)	0
dense_18 (Dense)	(None, 128)	32,896
dropout_9 (Dropout)	(None, 128)	0
dense_19 (Dense)	(None, 64)	8256
dropout_10 (Dropout)	(None, 64)	0
dense_20 (Dense)	(None, 32)	2080
dropout_11 (Dropout)	(None, 32)	0
dense_21 (Dense)	(None, 7)	231

Total params: 782,535
Trainable params: 782,535
Non-trainable params: 0

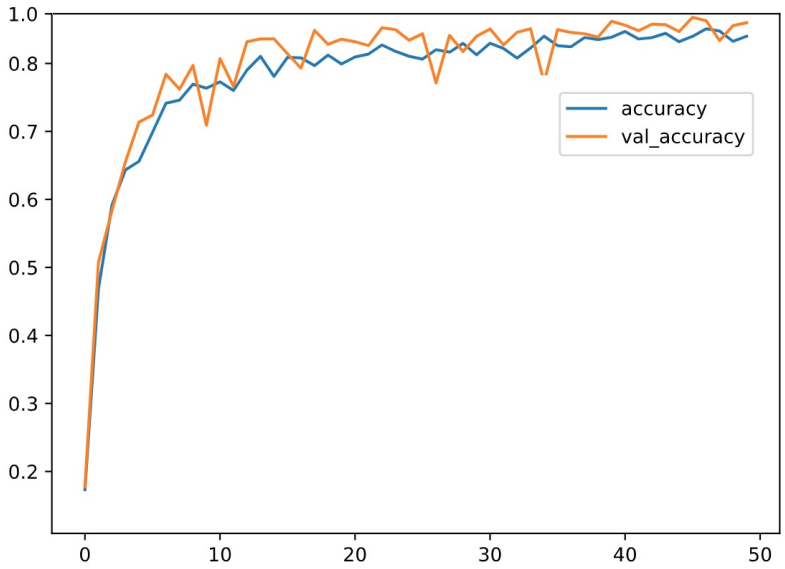
Figure 2. Architecture of the DNN Model.

3. Results

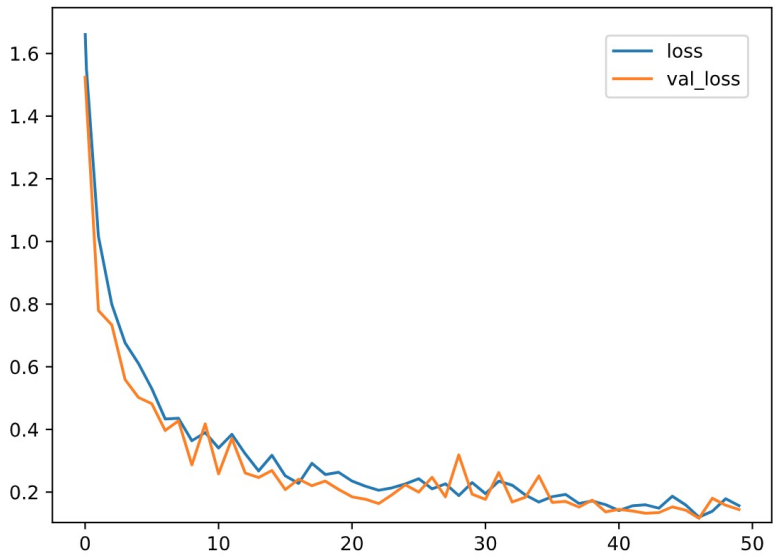
This section describes the results obtained from both the approaches over SER. Multiple evaluation metrics such as accuracy score, loss, classification report, and confusion matrix have been utilized throughout this research for evaluation. This section presents a detailed analysis of the performance of the previously described models on both datasets over multiple evaluation metrics.

3.1. Results on Approach I

The results obtained after training and testing the TESS dataset on the described DNN model are shown in Figure 3. Similar results have been found on the other datasets. The confusion matrices obtained after training and testing using the described DNN model on both the datasets are shown in Figure 4. Results depict high accuracy on both datasets. The confusion matrix shows more than 90% of samples are correctly classified for all the emotions. Confusion matrices in Figure 4 give a representation of the accuracies of the eight classes present in the RAVDESS dataset and seven classes present in the TESS dataset over DNN. Results show the outperformance over the TESS dataset with extremely low misclassifications.

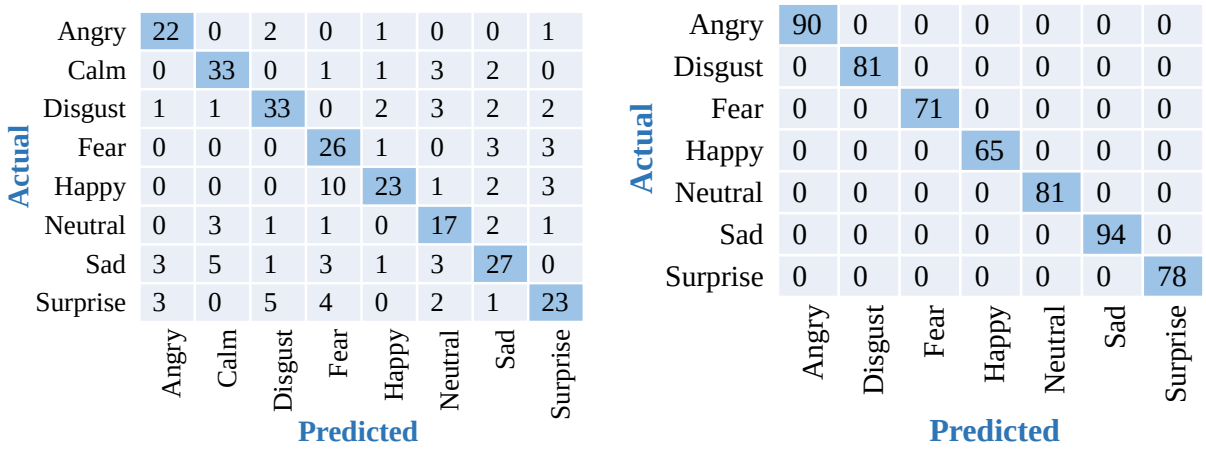


(**a**) Accuracy



(**b**) Loss

Figure 3. Accuracy and loss analyses of the proposed two-way feature extraction-based DNN on TESS dataset.

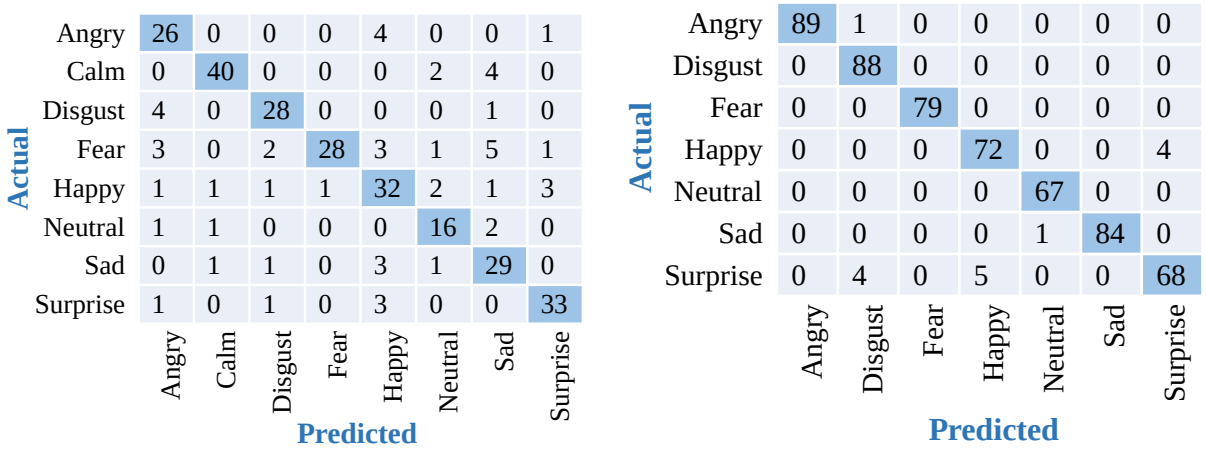


(**a**) RAVDESS (**b**) TESS

Figure 4. Confusion matrices for feature extraction and modeling using DNN over 2 datasets. Diagonal elements with dark blue color show accurately predicted classes.

3.2. Results on Approach II

The results obtained after training and testing of the RAVDESS dataset and the TESS dataset on the described VGG 16 transfer learning model are shown in Figure 5. Confusion matrices in Figure 5 represent the accuracies of the eight classes present in the RAVDESS dataset and seven classes present in the TESS dataset, obtained using VGG-16. This model gave us better accuracy than the DNN with low misclassifications. The results obtained are better than the previous approach. Due to the inbuilt feature extraction capability of VGG16 this model was able to outperform the other approach by giving an accuracy of 81.94% on the RAVDESS dataset and accuracy of 97.15% on TESS dataset.



(**a**) RAVDESS (**b**) TESS

Figure 5. Confusion matrices for feature extraction and modeling using VGG16 over 2 datasets. Diagonal elements with dark blue color show accurately predicted classes.

3.3. Comparison with State-of-Art Approaches

In this section, we have compared the results of the proposed approach with the existing state-of-the-art. Table 1 shows a comparative study with respect to research groups addressing speech emotion recognition using deep learning models.

Table 1. Comparative study.

SERIAL NO.	APPROACH	MODEL USED	DATASET USED	ACCURACY
01	Dissanayake [39]	CNN-LSTM (encoder)	RAVDESS	56.71%
02	Li et al. [40]	Multimodal Fine-Grained Learning	RAVDESS	74.7
03	Xu et al. [41]	Attention Networks	RAVDESS	77.4%
04	Proposed II Approach	2-D Feature Extration + VGG-16	RAVDESS	81.94

3.4. Comparative Analysis

To validate the effectiveness of the proposed approach, in addition to comparing results for two datasets, we have also compared results with the existing deep learning model of ResNet18. ResNet18 is an extremely popular CNN-based transfer learning model consisting of 18 layers. It has been loaded with the same weights and settings as VGG16. The results obtained after training and testing the RAVDESS dataset on the two datasets are shown in Figure 6. It shows the confusion matrices representing the best-performing models for the RAVDESS dataset. Out of the three models depicted in Figures 4–6, VGG-16 outperformed with highest accuracy and best classification over the RAVDESS dataset. Table 2 shows the accuracy scores for all the models over both the datasets. It depicts the highest and most stable accuracies obtained by the proposed approach utilizing the VGG16 model and 2-D feature extraction. ResNet-18 also showed comparatively high accuracy over TESS dataset. This result variation in the second case could be because in some cases ResNet-18 handles gradients differently than VGG-16.

(**a**) RAVDESS (**b**) TESS

Figure 6. Confusion matrices for feature extraction and modeling using Resnet-18 over 2 datasets. Diagonal elements with dark blue color show accurately predicted classes.

Table 2. Comparative analysis of RAVDESS and TESS.

SERIAL NO.	MODEL	RAVDESS	TESS
01	ResNet18	79.16%	96.26%
02	Proposed-I	73.95%	99.99%
03	Proposed-II	81.94%	97.15%

3.5. Comparison with Benchmark Algorithms

We have further compared the proposed approach with multiple deep learning and machine learning state-of-the-art models. For comparison, we have utilized four benchmark

models of Decision Tree Classifier, Random Forest Classifier, MLP Classifier, and ResNet18. All these classifiers have been utilized and set to their base configuration from sklearn. The model configuration of ResNet18 was set the same as VGG16 to keep effective comparison. The comparison of accuracy scores of the benchmark models and proposed models is shown in Table 3. It suggests the highest accuracies obtained by the proposed approaches with respect to other SOTA models.

Table 3. Comparative analysis of RAVDESS and TESS.

SERIAL NO.	MODEL	RAVDESS	TESS
01	Decision Tree	37.85%	3.21%
02	Random Forest	46.88%	7.68%
03	MLPClassifier	33.68%	15.54%
04	ResNet18	79.16%	96.26%
05	Proposed-I	73.95%	99.99%
06	Proposed-II	81.94%	97.15%

4. Conclusions

Identifying and processing human emotions via words is a challenging task. With the advent of machine learning and deep learning, several researchers have tried to address this. In the present work, a speech emotion recognition model has been proposed by using two-way feature extraction and deep transfer learning. Initially, two-way feature extraction has been proposed by utilizing the superconvergence to extract two sets of potential features from the speech data. Further, PCA is applied to the obtained first feature set. Thereafter, DNN with dense and dropout layers have been implemented on the important features obtained using PCA. On the other hand, a pre-trained VGG-16 model is applied to the second set of features to build the second model. Extensive experiments have been drawn and comparative analyses is performed in this work. Results revealed that the proposed models outperform the existing models in terms of various performance metrics. There are several limitations of this work, which can be the extension of this work in the future. The RAVDESS dataset consists only of North American speakers. Hence, the proposed approaches might give significantly less accuracy for people from different geographical areas. In future, we would like to apply the proposed model and other datasets as well. Similarly, this dataset takes into consideration people of median age. In future, we would like to extend this study to the vast vicinity characteristics of the subjects.

Author Contributions: Conceptualization, A.A. (Apeksha Aggarwal); methodology, A.S.; software and validation, A.A. (Ajay Agarwal), formal analysis and investigatio, N.C.; resources and data curation, D.S. and A.A.A.; visualization and supervision, A.A. (Aseel Alhadlaq) and H.-N.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MSIP) (NRF-2021R1A2B5B03002118) and this research was supported by the Ministry of Science and ICT (MSIT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-0-01835) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). The authors extend their appreciation to the Researchers supporting project number (RSP-2021/314) King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors extend their appreciation to the Researchers supporting project number (RSP-2021/314) King Saud University, Riyadh, Saudi Arabia.

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

References

1. Lech, M.; Stolar, M.; Best, C.; Bolia, R. Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Front. Comput. Sci.* **2020**, *2*, 14. [CrossRef]
2. 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]
3. Joy, J.; Kannan, A.; Ram, S.; Rama, S. Speech Emotion Recognition using Neural Network and MLP Classifier. *IJESC* **2020**, *2020*, 25170–25172.
4. Damodar, N.; Vani, H.; Anusuya, M. Voice emotion recognition using CNN and decision tree. *Int. J. Innov. Technol. Exp. Eng.* **2019**, *8*, 4245–4249.
5. Noroozi, F.; Sapiński, T.; Kamińska, D.; Anbarjafari, G. Vocal-based emotion recognition using random forests and decision tree. *Int. J. Speech Technol.* **2017**, *20*, 239–246. [CrossRef]
6. Eom, Y.; Bang, J. Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum Coefficients. *J. Inf. Commun. Converg. Eng.* **2021**, *19*, 148–154.
7. Rezaeipanah, A.; Mojarad, M. Modeling the Scheduling Problem in Cellular Manufacturing Systems Using Genetic Algorithm as an Efficient Meta-Heuristic Approach. *J. Artif. Intell. Technol.* **2021**, *1*, 228–234. [CrossRef]
8. Krishnamoorthi, R.; Joshi, S.; Almarzouki, H.Z.; Shukla, P.K.; Rizwan, A.; Kalpana, C.; Tiwari, B. A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *J. Healthc. Eng.* **2022**, *2022*, 1684017. [CrossRef]
9. Dubey, M.; Kumar, V.; Kaur, M.; Dao, T.P. A systematic review on harmony search algorithm: Theory, literature, and applications. *Math. Probl. Eng.* **2021**, *2021*, 5594267. [CrossRef]
10. Shukla, P.K.; Zakariah, M.; Hatamleh, W.A.; Tarazi, H.; Tiwari, B. AI-DRIVEN Novel Approach for Liver Cancer Screening and Prediction Using Cascaded Fully Convolutional Neural Network. *J. Healthc. Eng.* **2022**, *2022*, 4277436. [CrossRef] [PubMed]
11. Weiqiao, Z.; Yu, J.; Zou, Y. An experimental study of speech emotion recognition based on deep convolutional neural networks. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 827–831. [CrossRef]
12. Kurpukdee, N.; Kasuriya, S.; Chunwijitra, V.; Wutiwiwatchai, C.; Lamsrichan, P. A study of support vector machines for emotional speech recognition. In Proceedings of the 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), Chonburi, Thailand, 7–9 May 2017; pp. 1–6.
13. Shukla, P.K.; Shukla, P.K.; Sharma, P.; Rawat, P.; Samar, J.; Moriwal, R.; Kaur, M. Efficient prediction of drug–drug interaction using deep learning models. *IET Syst. Biol.* **2020**, *14*, 211–216. [CrossRef]
14. Liu, J.; Liu, Z.; Sun, C.; Zhuang, J. A Data Transmission Approach Based on Ant Colony Optimization and Threshold Proxy Re-encryption in WSNs. *J. Artif. Intell. Technol.* **2022**, *2*, 23–31. [CrossRef]
15. De Luca, G. A survey of NISQ era hybrid quantum-classical machine learning research. *J. Artif. Intell. Technol.* **2022**, *2*, 9–15. [CrossRef]
16. Sultana, S.; Iqbal, M.Z.; Selim, M.R.; Rashid, M.M.; Rahman, M.S. Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks. *IEEE Access* **2021**, *10*, 564–578. [CrossRef]
17. Lee, K.H.; Choi, H.K.; Jang, B.T. A study on speech emotion recognition using a deep neural network. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 16–18 October 2019; pp. 1162–1165.
18. Kaur, M.; Kumar, V. Parallel non-dominated sorting genetic algorithm-II-based image encryption technique. *Imaging Sci. J.* **2018**, *66*, 453–462. [CrossRef]
19. Pandey, S.; Shekhawat, H.; Prasanna, S. Deep Learning Techniques for Speech Emotion Recognition : A Review. In Proceedings of the 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16–18 April 2019. [CrossRef]
20. Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Sarma, K.K.; Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. *Interspeech* **2018**, *2018*, 3097–3101.
21. Li, P.; Song, Y.; McLoughlin, I.V.; Guo, W.; Dai, L.R. An attention pooling based representation learning method for speech emotion recognition. In Proceedings of the ISCA Conference, Los Angeles, California, USA, 2–6 June 2018.
22. Palo, H.; Mohanty, M.N.; Chandra, M. Use of different features for emotion recognition using MLP network. In *Computational Vision and Robotics*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 7–15.
23. 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. *arXiv* **2017**, arXiv:1706.00612.
24. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef] [PubMed]

25. Luo, D.; Zou, Y.; Huang, D. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. *Interspeech* **2018**, *2018*, 152–156.
26. 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.
27. 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.
28. Tao, F.; Liu, G. Advanced LSTM: A study about better time dependency modeling in 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. 2906–2910.
29. Lee, J.; Tashev, I. High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. *Interspeech* **2015**, *2015*, 336. [CrossRef]
30. Rokach, L.; Maimon, O., Decision Trees. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2005; pp. 165–192.
31. Ali, J.; Khan, R.; Ahmad, N.; Maqsood, I. Random forests and decision trees. *Int. J. Comput. Sci. Issues (IJCSI)* **2012**, *9*, 272.
32. Ramchoun, H.; Idrissi, M.A.J.; Ghanou, Y.; Ettaouil, M. Multilayer Perceptron: Architecture Optimization and Training. *Int. J. Interact. Multim. Artif. Intell.* **2016**, *4*, 26–30. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Lok, E.J. Toronto Emotional Speech Set (TESS). 2019. Available online: https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess (accessed on 16 December 2021).
35. Livingstone, S.R. RAVDESS Emotional Speech Audio Emotional Speech Dataset. 2018. Available online: https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio (accessed on 6 December 2021).
36. Satapathy, S.; Loganathan, D.; Kondaveeti, H. K.; Rath, R. Performance analysis of machine learning algorithms on automated sleep staging feature sets. *CAAI Trans. Intell. Technol.* **2021**, *6*, 155–174. [CrossRef]
37. Zou, Q.; Xiong, K.; Fang, Q.; Jiang, B. Deep imitation reinforcement learning for self-driving by vision. *CAAI Trans. Intell. Technol.* **2021**, *6*, 493–503. [CrossRef]
38. Chen, R; Pu, D; Tong, Y; Wu, M. Image-denoising algorithm based on improved K-singular value decomposition and atom optimization *CAAI Trans. Intell. Technol.* **2022**, *7*, 117–127.
39. Dissanayake, V.; Zhang, H.; Billinghurst, M.; Nanayakkara, S. Speech Emotion Recognition'in the Wild'Using an Autoencoder. *Interspeech* **2020**, *2020*, 526–530.
40. Li, H.; Ding, W.; Wu, Z.; Liu, Z. Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition. *arXiv* **2020**, arXiv:2010.12733.
41. Xu, M.; Zhang, F.; Zhang, W. Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access* **2021**, *9*, 74539–74549. [CrossRef]