Proceedings of the 5th International Conference on Smart Systems and Inventive Technology (ICSSIT 2023)
IEEE Xplore Part Number: CFP23P17-ART; ISBN: 978-1-6654-7467-2

Deep Learning Model for Emotion Prediction from Speech, Facial Expression and Videos

Chepuri Rajyalakshmi
PG Scholar
Dept. of Computer Science and Engineering
Narasaraopeta Engineering College
(A), Andhra Pradesh, India

K.LakshmiNadh
Professor
Dept. of Computer Science and Engineering
Narasaraopeta Engineering College
(A), Andhra Pradesh, India

M Sathyam Reddy
Asst. Professor
Dept. of Computer Science and Engineering
Narasaraopeta Engineering College
(A), Andhra Pradesh, India

2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) | 978-1-6654-7467-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSIT55814.2023.10061135

***Abstract*— The rapid development of computer vision and machine learning in recent years has led to fruitful accomplishments in a variety of tasks, including the classification of objects, the identification of actions, and the recognition of faces, among other things. Nevertheless, identifying human emotions remains one of the most difficult tasks to do. To find a solution to this issue, a significant amount of work has been put in. In order to achieve higher accuracy in this reactivity towards a variety of speeches and vocal-based methods, computer intelligence, natural language modelling systems, and other similar technologies have been used. The examination of the emotions has the potential to be useful in a number of different settings. Cooperation with human computers is one example of such a field. Computers can help customers recognize emotions, make wiser decisions, and create more lifelike human-robot interactions. In recent times, there has been a lot of focus placed on the ability to forecast dynamic facial emotion expressions in videos. Therefore, this work proposes a deep convolutional neural networks (CNNs) model for emotion prediction from speech samples, facial expression images, and videos with enhanced prediction accuracy and reduced loss. In addition, the speech CNN model also utilizes mel-frequency Cepstrum coefficients (MFCC) as feature extraction from given speech samples. The proposed MFCC-CNN model resulted in superior performance than traditional models.**

Keywords*— *Speech emotion, facial emotion, convolutional neural network, Mel Frequency Cepstral Coefficient, speech emotion recognition

I. INTRODUCTION

Over the course of the last several decades, facial emotion detection has received a significant amount of interest due to its applications in the cognitive sciences and affective computing. The researchers Ekman and colleagues recognized six fundamental facial expressions as fundamental emotional emotions that are universal and ubiquitous among human beings [1]. Human emotions are notoriously difficult to decipher, the development of emotion identification systems has become an integral part of human-computer interaction. There have been many different ways suggested for automated identification of face expressions [2]. The vast majority of previous, shallow techniques have concentrated their efforts on the independent analysis of static pictures, completely neglecting the temporal relationships between sequence frames in films. However, this temporal information is necessary in order to follow the subtle changes that occur in a person's face as they are expressing different emotions [3]. Recently, with the rise of deep learning, more encouraging findings have been published, when addressing the issue of automated face emotion identification using both geometric and photometric information.

Automatic facial expression recognition entails three stages: face recognition; feature extraction and categorization depending on how they are utilised to transmit emotions and input; and lastly, emotion identification based on the extracted features. While human user interfaces have come a long way since the days of the mouse and keyboard with innovations like automatic voice recognition and specialised interfaces for people with disabilities [4], they still fail to fully account for the entire range of people's interactive skills. When robots are able to read such emotional indicators, they will be able to provide users with more tailored support that meets their specific needs [5]. It is widely accepted, based on research conducted in the field of psychological science, that human emotions may be broken down into six archetypal feelings: shock, fear, disgust, fury, joy, and sorrow. When it comes to conveying certain emotions, the tone of one's voice and the look on one's face are both very important factors. Interpreting people's feelings has been an important area of study in recent years because of the information it may bring for a wide variety of applications [6]. People express their emotions, either deliberately or unconsciously, by the words they choose to speak and the facial expressions they make. Knowledge of many various kinds, such as verbal, written, and visual information, may be used in the process of emotion interpretation. Since ancient times, speech and facial expression have been a helpful tool for understanding sentiments. Additionally, they have disclosed a variety of aspects, including mentality, which is one of those aspects [7]. The task of determining the emotions that lie underlying these remarks and facial expressions is a massive one that is fraught with difficulty. In order to discover a solution to this problem, researchers from a wide variety of scientific fields are combing through data from a variety of sources, such as voice and facial expressions, in search of a more accurate method of identifying human emotions [8]. In order to achieve higher accuracy in this reactivity towards a variety of speeches and vocal-based methods, computer intelligence, natural language modelling systems, and other similar technologies have been used. The examination of the emotions has the potential to be useful in a number of different settings. Cooperation with human computers is one example of such a field. Computers are capable of making more enlightened decisions, assisting customers in recognizing their feelings, and contributing to the creation of more lifelike human-robot interactions.

II. LITERATURE SURVEY

Over the course of the last several decades, research on facial emotion recognition (FER) systems has been getting a lot of interest because to the fast growth of methods used in artificial intelligence [9]. Several different feature-based strategies have been researched for use with FER systems. These methods locate a face area inside a picture and then extract geometric or physical information from that region. The geometric aspects often comprise the connection between face components. Instances of characteristic instances of geometric characteristics include facial landmark points [10]. The characteristics of the global face areas or the various sorts of information on the facial regions may be retrieved and used as appearance features [11]. Principal component analysis, local binary pattern histograms, and a few more types of analyses are often included in the global futures. Several of the investigations broke the face area up into smaller, more localized sections and then retrieved characteristics of appearance that were unique to each section [12]. First, from among these local areas, the regions with the most significance are identified, which ultimately leads to an improvement in the identification accuracy. In particular, the Artificial neural network (ANN) [13] has performed very well in a number of experiments, including those involving object identification, and FER. Despite the fact that deep-learning-based algorithms have obtained better outcomes than traditional methods.

Speech signals are among the most natural means of human communication, and they also have the advantage of being straightforward and easy to measure in real time. The linguistic content of speech signals is accompanied by implicit paralinguistic information about the speakers, which may include their feelings [14]. The majority of speech-emotion recognition algorithms, such as FER, extract acoustic features. Consequently, integrating the proper aural characteristics is essential. Numerous researches have provided evidence supporting a link between emotional voices and acoustic aspects of the voice. Because of this, determining the best collection of features to use in speech-emotion recognition is an essential step in the process [15].

When a machine is trying to infer human emotions, using voice signals and face pictures may be beneficial for achieving recognition that is both accurate and natural. In order to accomplish this goal, the information on the feelings must be mixed in a suitable manner and to varying degrees. The vast majority of multimodal research concentrate on three different strategies: combining features, fusing decisions, and concatenating models [16]. The technique of deep learning, which is used in a variety of sectors, may play an important part in the process of combining many inputs. Model concatenation is a straightforward technique that may be used when there is a need to combine models that take a variety of inputs. Each encoded tensor is produced by a model that takes a unique set of data as its input. Using the concatenate function, it is possible to link the tensors that are included inside each model. In [17] authors considered voice signals and transformed them into mel-spectrogram pictures such that a 2D-CNN could use the image as an input. This was done after they had concatenated the two networks. The goal of decision fusion is to re-distinguish between candidates by processing the category that is produced by each model and using the unique criteria [18]. In order to do this, the SoftMax functions of the various kinds of networks are combined via the process of computing the dot product with weights, where the total of all the weights equals.

They integrated the CNN and recurrent neural networks (RNN) techniques [19] in order to transform voice signals into features, and they made use of the MFCC. The weighted-decision fusion approach was utilized by these researchers in order to combine the speech signals with facial emotions. Deep temporal appearance networks and deep temporal geometry networks were the two varieties of deep networks, it has been pre-trained [20]. Because current approaches rely mostly on shallow fusion, it is necessary to develop a more comprehensive fusion model.

III. PROPOSED METHOD

When it comes to communicating with other humans, emotions are an absolutely necessary component. Emotions, actions, and ideas are intricately connected to one another in such a manner that the interplay of these three factors determines how we conduct ourselves and the choices we make. Because of this, there has been an increasing interest in this particular field of scientific inquiry throughout the course of the previous several years. The automatic identification of emotions has a variety of applications that may be used to improve those feelings. For instance, human–computer interaction, since determining the emotional state of a person who uses a computer system would make it possible to generate a more natural, productive, and intelligent relationship between the two parties. An additional domain is human-to-human interaction monitoring, which is useful since it enables the detection of hostile or unwelcome scenarios. This research discusses the automated identification of emotions based not just on facial expressions but also on voice and videos. The approach that was presented made use of deep learning CNN methods such as the production of corpora, the selection of features, the building of an acceptable classification scheme, and the merging of this information with other sources of information such as text.

Figure 1 shows the proposed block diagram of face and speech-based emotion recognition. RACVDESS dataset is considered to implement this work, which contains both speech and face data files. Then, pre-processing operation is carried out on both datasets performed, which removed the noises from facial images and speech files. Then, MFCC features are extracted only from speech data. Then, CNN model is trained with the both speeches based MFCC features and pre-processed facial data. Finally, test face and speech data are applied and test features are compared with the pre-trained CNN model features. Finally, the predicted emotion is obtained through this AI-CNN model from both face and speech data.

A. Dataset description

For the face emotion identification model, we utilized 28,709 photos containing seven distinct emotions, including fear, disgust, surprise, anger, and happiness. For the purpose of developing a speech emotion recognition model, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset was used. The data rate, sampling frequency, and format of spoken audio-only files generated by the RAVDESS are 16 bits, 48 kilohertz, and.wav, respectively. This section of RAVDESS is home to 1440 different files, including: 60 trials per actor × 24 actors = 1440. The RAVDESS is made up of 24 professional actors,

12 of whom are female and 12 of whom are male. These actors speak in a neutral North American dialect while pronouncing two lexically matched phrases. Expressions of calm, happiness, sadness, anger, fear, surprise, and disgust are all included in the category of spoken emotions. Each expression may be created at two different degrees of emotional intensity, normal and strong, as well as a third expression that is neutral.

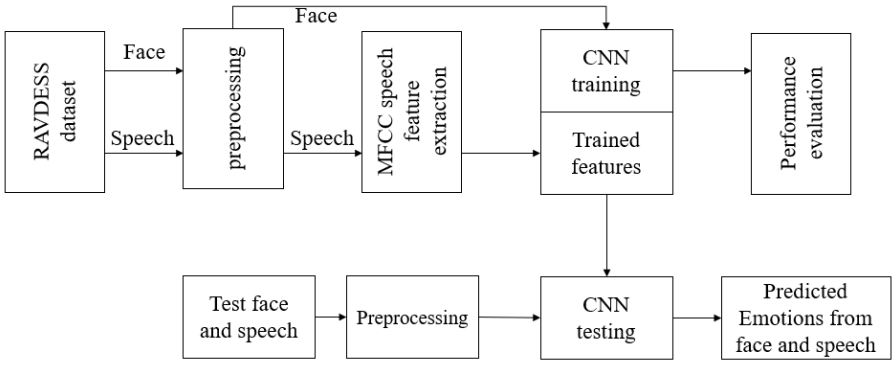

Fig. 1. Proposed block diagram.

B. Preprocessing

The dataset contains noises, missing values, which caused to complicated training of CNN model. Further, it will reduce the classification, prediction performance. So, the data preprocessing operation is performed to overcome these problems. The preprocessing operation will replace unknown symbols, missing vales with the known nearest values. Splitting the Dataset: Our dataset is divided into two distinct categories, the training set and the test set during the preprocessing phase of CNN. These categories are named respectively the training set and the test set. This is one of the most significant activities since enhancing the overall performance of deep learning CNN model is one of the keys aims of the processes for data preparation, and this is one of the chores that has to be completed. Take into consideration the following possibility: After training CNN model on a certain dataset, we tested it on a completely other dataset to see how well it performed. If this occurs, our model will have a more difficult time understanding the links that exist among the various models. If we train our model incredibly well, and if its training accuracy is also fairly excellent, but then we feed it with a new dataset, then the performance will fall; however, this only happens if we train it exceptionally well. As a result of this, our objective while constructing a model for CNN is to make certain that it performs well not just with the training set but also with the test dataset.

C. MFCC Feature extraction

The MFCC is a well-known unsupervised learning technique that may decrease the dimensionality of data in many ways. It increases the material's interpretability and decreases the quantity of information that is lost. It makes the data easy to plot in both two and three dimensions and aids in identifying the most significant parts of a dataset. The MFCC is beneficial for discovering a sequence of linear combinations of the researched variables.

D. CNN Models

In addition to this, it demonstrates how the data travels through the system and how its state is altered as a result of a number of changes. Figure 2 shows the proposed deep CNN model for emotion detection using speech recognition. Figure 3 shows the proposed deep CNN model for emotion detection from facial expressions. Finally, the combination of face and speech is considered as the video-based emotion recognition.

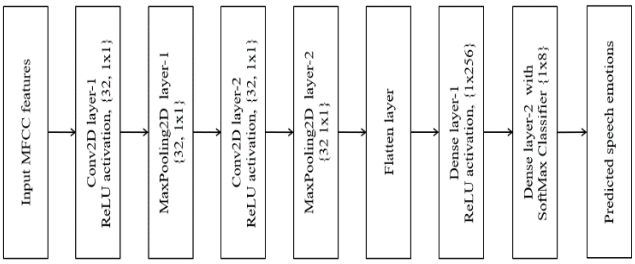

Fig. 2: Proposed deep CNN model for emotion detection using speech recognition.

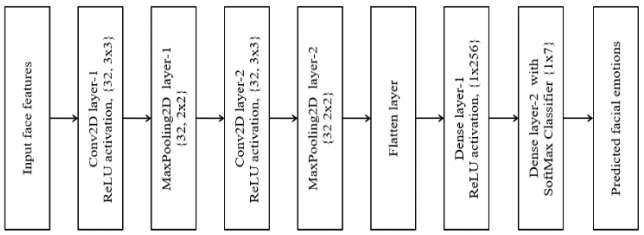

Fig. 3: Proposed deep CNN model for emotion detection from facial expressions.

It is a method of visual representation that illustrates the flow of information and the changes that take place when data is moved from input to output. A system may be represented by it at any degree of abstraction, and it can be partitioned into layers that reflect increasing information flow and functional complexity. In addition, it can be used to describe a system in any way. Normalize the data: Before completing the principal component analysis, normalize the data. This will guarantee that each characteristic has a mean value of zero and a variance value of one. Constructing a square matrix to describe the connection between two or more characteristics in a multidimensional dataset is what is meant by "building the covariance matrix." Find the Eigenvectors and Eigenvalues: Perform the calculations necessary to determine the eigenvalues and eigenvectors/unit vectors. Scalars known as eigenvalues are used to multiply

the covariance matrix's eigenvector in order to get the variance matrix. Determine the number of primary components after sorting the eigenvectors in descending order from highest to lowest. Finally, SoftMax classifier identifies the different emotions based on MFCC extracted features.

IV. Results And Discussions

This section gives the detailed results of simulation analysis. Figure 4 illustrate the sample test images of emotion prediction from given facial expressions, where it includes all the emotions such as sad, angry, neutral, disgusted, surprised, and fearful. Figure 5, and Figure 6 discloses the obtained prediction accuracy and loss performance using proposed deep CNN from facial expression, speech, and videos. From both the figures, it is observed that proposed deep CNN obtained superior performance for emotion prediction from videos as compared to both facial expression and speech inputs. Table 1 compares the performance of proposed CNN models performance with existing approaches. Here, the proposed deep CNN resulted in superior performance as compared to basic ANN, RNN, and LSTM models.

Fig. 4: Sample test images of emotion prediction.

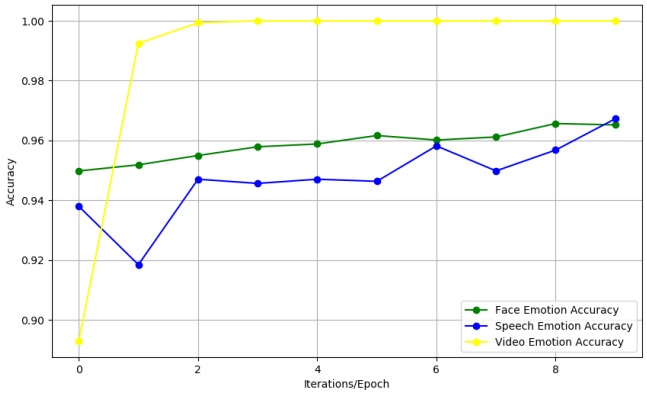

Fig. 5: Performance comparison of prediction accuracy using proposed deep CNN with speech, facial expression, and videos.

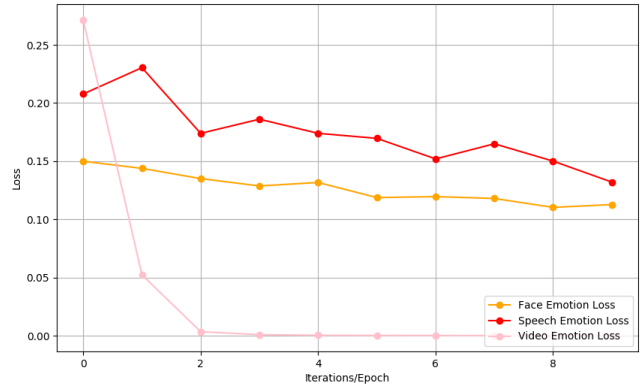

Fig. 6: Performance comparison of prediction loss using proposed deep CNN with speech, facial expression and videos.

Table 1. Accuracy performance comparison

Dataset	ANN [13]	RNN [15]	LSTM [17]	Proposed Deep CNN
Facial expression	76.26%	81.23%	90.345%	96%
Speech expression	87.34%	89.34%	92.345%	97%
Video expression	91.26%	93.45%	95.78%	100%

V. Conclusion

Interpreting people's feelings has been an important area of study in recent years because of the information it may bring for a wide variety of applications. People express their emotions, either deliberately or unconsciously, by the words they choose to speak and the facial expressions they make. Knowledge of many various kinds, such as verbal, written, and visual information, may be used in the process of emotion interpretation. Therefore, this work proposed a deep CNN model for emotion prediction from speech, facial expression, and videos with enhanced prediction accuracy and reduced loss. In addition, the speech CNN model utilized MFCC as feature extraction from given speech samples. This work can be extended with other facial expression for improved performance.

REFERENCES

[1]. Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." *Sensors* 21.4 (2021): 1249.

[2]. Kwon, Soonil. "Optimal feature selection based speech emotion recognition using two- stream deep convolutional neural network." *International Journal of Intelligent Systems* 36, no. 9 (2021): 5116-5135.

[3]. Avots, E., Sapiński, T., Bachmann, M., & Kamińska, D. (2019). Audiovisual emotion recognition in wild. Machine Vision and Applications, 30(5), 975-985.

[4]. Wang, Xusheng, Xing Chen, and Congjun Cao. "Human emotion recognition by optimally fusing facial expression and speech feature." *Signal Processing: Image Communication* 84 (2020): 115831.

[5]. Kerkeni, Leila, et al. "Automatic speech emotion recognition using machine learning." *Social media and machine learning*. IntechOpen, 2019.

[6]. Pandey, Sandeep Kumar, Hanumant Singh Shekhawat, and SR Mahadeva Prasanna. "Deep learning techniques for speech emotion recognition: A review." *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2019.

[7]. Özseven, Turgut. "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition." *Applied Acoustics* 142 (2018): 70-77.

[8]. Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2018.

[9]. Schoneveld, Liam, Alice Othmani, and Hazem Abdelkawy. "Leveraging recent advances in deep learning for audio-visual emotion recognition." *Pattern Recognition Letters* 146 (2021): 1-7.

[10]. Vryzas, Nikolaos, et al. "Continuous speech emotion recognition with convolutional neural networks." *Journal of the Audio Engineering Society* 68.1/2 (2020): 14-24.

[11]. Neumann, Michael. "Cross-lingual and multilingual speech emotion recognition on english and french." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[12]. Hossain, M. Shamim, and Ghulam Muhammad. "Emotion recognition using deep learning approach from audio–visual emotional big data." *Information Fusion* 49 (2019): 69-78.

[13]. Pan, Zexu, et al. "Multi-modal attention for speech emotion recognition." *arXiv preprint arXiv:2009.04107* (2020).

[14]. Jain, Neha, et al. "Hybrid deep neural networks for face emotion recognition." *Pattern Recognition Letters* 115 (2018): 101-106.

[15]. Jannat, Rahatul, et al. "Ubiquitous emotion recognition using audio and video data." *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 2018.

[16]. Dai, Dongyang, et al. "Learning discriminative features from spectrograms using center loss for speech emotion recognition." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[17]. Mittal, Trisha, et al. "Emoticon: Context-aware multimodal emotion recognition using frege's principle." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[18]. Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." *IEEE Access* 8 (2020): 79861-79875.

[19]. Grundmann, Felix, Kai Epstude, and Susanne Scheibe. "Face masks reduce emotion-recognition accuracy and perceived closeness." *Plos one* 16.4 (2021): e0249792.

[20]. Latif, Siddique, et al. "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition." *IEEE Transactions on Affective computing* 13.2 (2020): 992-1004.