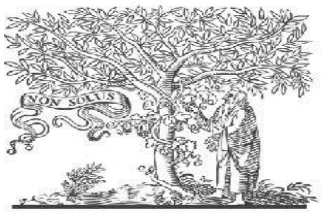
International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT



ELSEVIER
SSRN

IJIEMR Transactions, online available on 13th Apr 2022. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=ISSUE-04

DOI: 10.48047/IJIEMR/V11/I04/24

Title **Speech Emotion Recognition (SER) Using Convolutional Neural Networks (CNN)**

Volume 11, Issue 04, Pages: 152-159

Paper Authors
B. Rishitha, Ch. Kartheek Naidu, Ch. Sowmya, Ch. Chaitanya Jyothi,
Mrs. K. Divya

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Speech Emotion Recognition (SER) Using Convolutional Neural Networks (CNN)

B. Rishitha[1], Ch. Kartheek Naidu[2], Ch. Sowmya[3], Ch. Chaitanya Jyothi[4], Mrs. K. Divya[5]

[1], Student, Dept of Computer Science and Engineering, SRGEC, Gudlavalleru-521356, India
[2], Student, Dept of Computer Science and Engineering, SRGEC, Gudlavalleru-521356, India
[3], Student, Dept of Computer Science and Engineering, SRGEC, Gudlavalleru-521356, India
[4], Student, Dept of Computer Science and Engineering, SRGEC, Gudlavalleru-521356, India
[5], Assistant Professor, Dept of AI&DS, SRGEC, Gudlavalleru-521356, India
Email:batturishitharamaraju@gmail.com,kartheeknaidu2001@gmail.com,sowmyachinnam21@gmail.com, chaitanyajyothi246@gmail.com, divyakothapalli137@gmail.com

Abstract:

The Automated Speech Emotion Recognition process is difficult due to the disparity between acoustic features and human emotions, which is heavily dependent on the discriminative acoustic characteristics collected for a given recognition job. Different people have different emotions and exhibit them in different ways. When contemplating various subjects, the intensity of speech emotion varies, and pitch fluctuations are emphasized. As a result, voice emotion recognition is a difficult task in computer vision. A variety of strategies have been employed to extract emotions from signals in the literature on speech emotion recognition (SER), including well-known speech analysis and classification approaches, but each model has significant disadvantages. The old standard technique, which was implemented using a Support Vector Machine (SVM) classifier, was employed to identify emotions. The SVM classifier achieved an 81 percent recognition rate. And the other systems, while using hybrid classifiers to increase complexity, can only distinguish four emotions. Our suggested system's main goal is to build voice emotion recognition based on the Convolutional Neural Network (CNN) algorithm using the RAVDESS dataset, which uses multiple modules for emotion detection and classifiers to identify emotions such as happiness, surprise, anger, neutral mood, sadness, and so on. The voice samples serve as the dataset for the speech emotion detection system, and the characteristics are retrieved from these speech samples using

Tensor flow tools. Our ultimate aim is to recognize the emotion in a speech signal with higher accuracy than currently available methods.

Keywords: Speech Emotion, recognition rate, Tensor flow, Support Vector Machine(SVM), Convolutional Neural Network(CNN)

1. Introduction

Speech recognition as a means of recognizing human thinking is a cutting-edge hotspot in human-computer interaction. Human feelings are one-of-a-kind intellectual experiences that could be supported together in a range of methods, including through body language, voice, and body language. Speech is the most powerful and valuable of these modalities, which is why scholars are interested in working in the field of speech processing. Speech emotion awareness (SER) has a wide range of real-world applications, including fitness and medicine, recognizing caller irritation and disappointment in name middle services, lie detection, higher tutoring systems, and shrewd assistance [1], [2], [3], and so on. Speakers, languages, and cultures all have emotional subjectivity exclusive recording stipulations, SER is a tough issue [4]. In two processes, human feelings are extracted from the speech: characteristic extraction and classification. Pitch, intensity, formants, and length were initially used to provide emotional significance clues. In the past, Fourier parameter models have been used for state-of-the-art [5] [6] advances. All components of speech perception are influenced by speaker recognition, acoustic segmentation, and speech signals. Numerous other combinations of facets, such as ISO9-emotion and GeMaps [7], [8,] are also mentioned in the literature. These characteristics are usually prosodies, with statistically significant values such as mean, minimum, maximum, and general deviation. Typical classifiers such as GMM, HMM, SVM, and others are used to categorize the amazing representations of voice signals. Although a significant amount of research has been done in this sector as far as handcrafted points are concerned, there is still uncertainty about which kind of characteristics are appropriate for which class of thoughts [9], [10]. Expertise and a thorough understanding of the data are also required for the extraction of more appropriate aspects. Deep learning has completely transformed these tactics. Numerous scientists have achieved incredible developments in their findings. However, it's difficult to encode the last output into the equivalent enter samples [11]. Deep learning is typically identified using one of two ways. The first is to supply raw data to a model so that it can locate relevant elements, and the second is to employ comprehensive data visualization as input. Sequential patterns, such as recurrent neural networks, were originally utilized in audio statistics processing to master temporal information. However, Convolutional neural networks have been integrated with these architectures for richer secular data and a higher level of difficulty with various manageable frameworks [12], [13].

II Literature Survey

Emotion evaluation, widely called Information Extraction as well as Emotion Cognitive Computing, is a rapidly emerging topic that combines Natural Language Processing, textual content evaluation, and the characterization and learning about mental responses from a set of statistics or textual content information. It is still a growing field in the ground of unstructured textual

extraction. Many businesses use sentiment analysis for product evaluations, and social media comments, and a small percentage of the time, it's used to determine whether or not textual material is favorable, negative, or neutral. During this study, developers plan to use the most advanced Python language to perform rule-based total processes that define a list of norms as well as components such as Quintessential Text Analysis methodologies, separating, smart contracts, area of strategy is applied, and laptop learning parsing for sentiment evaluation.

Emotion detection is a critical and challenging issue, and function extraction has a considerable impact on SER performance. According to developments in deep learning, scientists can now focus on the end-to-end shape and validate a beautiful and successful solution. Therefore in the article, designers start introducing the ADRNN (distended Classifier to dense block but rather BILSTM predicated mostly on interest methodology) as a new structure for practicing speech emotion attention, that can also carry the advantage of the strengths of various channels while overcoming the drawbacks of using them by itself, but instead, we evaluate it using the well-known Uniqueness repository and the Berlin EMODB codification. Using a compressed CNN rather than a pooling layer can help the mannequin accumulate more receptive fields. The pass-by connection can then store additional ancient data according to the uppermost stratum, The BILSTM level, on the other hand, is utilized to look into lengthy connections based on the found nearby characteristics. We also apply the interesting technique to improve similar speech feature extraction. In addition, they modify its damage characteristic to detect the trained model in conjunction with the input impedance, which leads to better classification results. They extract the 3D Log-Mel frequency values from fresh alerts as well as feed those into their suggested technique as sentimental discussions were also modified of such graphical representations, achieving incredible overall effectiveness of 74.96% intermediate accuracies in the voice experiment and 69.32% approximate precision inside the person speaking investigation. That's greater than the 64.74% in the IEMOCAP application's spontaneous emotional speech from earlier strategies. Furthermore, using the Berlin EMODB of speaker-dependent and speaker-independent scans, we recommend the networks that achieve awareness accuracies of 90.78 percent and 85.39 percent, accordingly, which are either higher than the 88.30% and 82.82% obtained by earlier work. They also do a cross-corpus scan between the aforesaid databases to validate the robustness and generalization, and we get the preferred 63.84 percent attention accuracy in the end.

III Proposed System

In this paper, we propose employing a deep mastering technique on Log-Mel-spectrograms of fragmented audio statements to recognize speech emotion using Deep CNN (convolutional neural network). The proposed architecture is used to extract emotion-related characteristics from a dataset. We were able to distinguish emotion-based speech using this dataset. The proposed model's performance is demonstrated by several rigorous experiments on various datasets, and the results are differentiated to contemporary CNN architectures. The suggested CNN network achieves 95% accuracy for eight emotional classifications.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

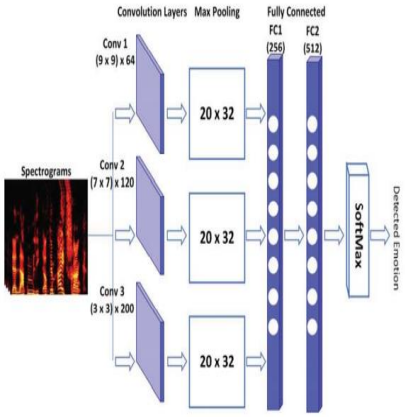


Fig.1: Proposed CNN architecture for SER

The phases of the proposed system are as follows:

1. Upload Emotion Dataset: Using this module we can upload speech emotion audio files of several actors using the RAVDESS dataset and read all audio records and convert all speech audio files into train and test array. After uploading the dataset, it will be preprocessed and details about the dataset will be displayed.

2. Train Dataset Using CNN: Using this module all train and test arrays will be passed to CNN layers and then these layers will filter all speech audio files array and then build a classification model where each audio file will be classified into one of eight types such as 'Angry', 'Sad', 'Fearful', 'surprise' 'Happy' etc.,

3. Predict Speech Emotion: Using this module we will upload a test speech and then CNN will classify the emotion of the speech.

3.1 CNN Algorithm

CNN (Convolutional Neural Networks) is a sort of human brain that can be used to address a wide range of issues. Without a doubt, it is the most popular deep learning architecture. The great popularity and efficacy of content have spurred the recent surge in interest in deep learning. AlexNet piqued CNN's interest in 2012, and it has grown tremendously since then. Researchers progressed from 8 layers of AlexNet to 152 levels of ResNet in just three years.

CNN is presently the model of choice for any image-related problem. When it comes to precision, they outperform the competition. It's also utilized in things like recommender systems and natural language processing CNN's great element across its competitors is whether it detects required features autonomously, without the use of user participation. Given a large collection of pictures of cats and dogs, it learns distinctive features for each class on its own.

Furthermore, CNN is a computationally efficient algorithm. It uses specific convolution and pooling methods, as well as parameter sharing. CNN models may now operate on any device, making them more accessible to a wider audience.

Ultimately, this appears to be pure magic. We're working with an extremely powerful and efficient model that conducts autonomous feature extraction with superhuman precision (yes CNN models now do image classification better than humans). Hopefully, this paper will assist us in unraveling the mysteries of this amazing approach.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

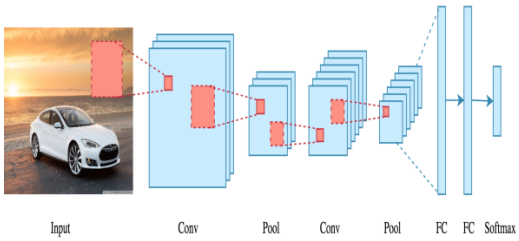


Fig 2 : CNN Structure

3.2 DATASET

Import libraries and define functions for charting the data using matplotlib to start this exploratory research. Not all plots will be made, depending on the data. We have downloaded this 'RAVDESS' dataset from Kaggle is one of the best websites for datasets. The dataset contains 1435 records of various emotional characteristics that have been categorized into actors.

Preprocessing and Training the model (CNN): The dataset is preprocessed such as audio reshaping, resizing, and conversion to an array form. Similar processing is additionally completed on the tested audio. A dataset of around eight different speech emotions is obtained, out of which any audio can be used as test audio for the software.

The CNN model is trained to recognize the test audio and the emotion it contains using the learned dataset. When the model has been effectively trained, the software can recognize the speech emotion Classification audio in the dataset. The test audio and trained model are compared to forecast the speech emotion after successful training and preprocessing.

By using this RAVDESS dataset here we have implemented this concept.

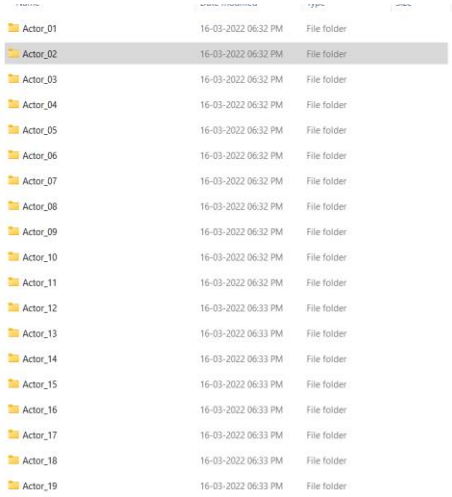


Fig 3. RAVDESS Dataset Information

IV Results

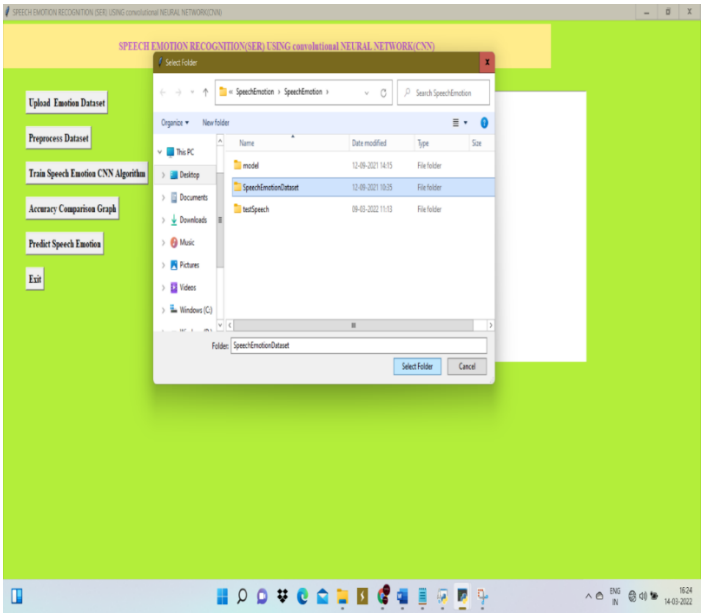


Fig 4: In the above screen we are uploading the dataset

Click on speech emotion dataset folder (RAVDESS dataset) then click on 'Select Folder' to upload all files and after uploading the file will get below screen

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig 5: dataset loaded successfully and we got total files details and types of emotion details

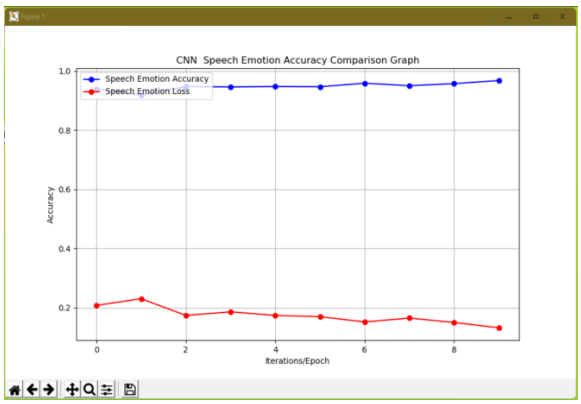


Fig 6: Accuracy Comparison Graph

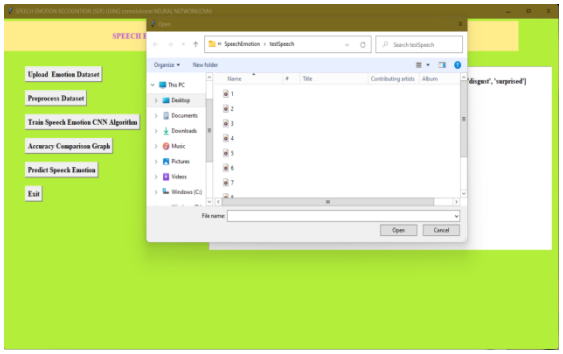


Fig 7: In the above screen we are uploading a speech file. based on speech we will get emotion

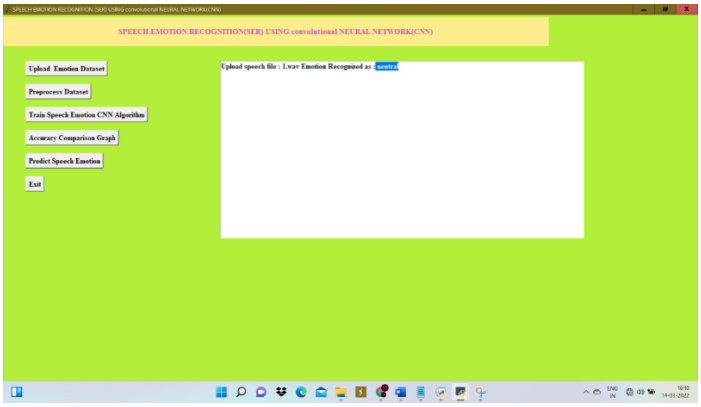


Fig 8: In the above screen we got emotional details based on the input file. Here we recognized emotion as neutral

5. Conclusion

Machine learning is a relatively new topic, however, it is quickly growing as a result of new worth checking for computer vision challenges motivated by industrial cases. Identifying emotions from speech, on the other hand, remains a difficult undertaking. On datasets, we employed an easy deep CNN structure proposed in this research for voice emotion classification. Furthermore, We discovered that CNN-based time-distributed systems score higher. This outcome comparability can be utilized as a starting point for future research, and we believe that using more concatenated CNNs will yield superior results. We intend to investigate the audio/video-based emotional speech identification task in the future.

References

1. Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 145-152.

2. Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor. "Emotion recognition in human-computer interaction." IEEE Signal processing magazine 18, no. 1 (2001): 32-80.

3. Kotti, Margarita, and Fabio Paterno. "speaker-independent emotion ` recognition exploiting a psychologically-inspired binary cascade classification schema." International journal of speech technology 15, no. 2 (2012): 131-150.

4. Minker, Wolfgang, Johannes Pittermann, Angela Pittermann, PetraMaria Strauß, and Dirk Buhler. "Challenges in speech-based hu- ¨ man-computer interfaces." International Journal of Speech Technology 10, no. 2 (2007): 109-119.

5. Busso, Carlos, Sungbok Lee, and Shrikanth Narayanan. "Analysis of emotionally salient aspects of fundamental frequency for emotion detection." IEEE transactions on audio, speech, and language processing 17, no. 4 (2009): 582-596.

6. Chauhan, Krishna, Kamlesh Kumar Sharma, and Tarun Varma. "Improved Speech Emotion Recognition Using Modified Mean Cepstral Features." In 2020 IEEE 17th India Council International Conference (INDICON), pp. 1-6. IEEE, 2020.

7. Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE transactions on acoustics, speech, and signal processing 28, no. 4 (1980): 357-366.

8. Eyben, Florian, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, ¨ Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, et al. "The Geneva ´ minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." IEEE transactions on affective computing 7, no. 2 (2015): 190-202.

9. El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern recognition 44, no. 3 (2011): 572-587.

10. Guizzo, Eric, Tillman Weyde, and Jack Barnett Leveson. "Multi-timescale convolution for emotion recognition from speech audio signals." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6489-6493.

1. Lipton, Zachary C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16, no. 3 (2018): 31-57.

12. Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." Biomedical Signal Processing and Control 47 (2019): 312-323. IEEE, 2020.

13. Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. "Automatic speech emotion recognition using recurrent neural networks with local attention." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227-2231. IEEE, 2017.

14. Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional

neural networks." In 2015 international conference on affective computing and intelligent interaction (ASCII), pp. 827-831. IEEE, 2015.

15. Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Learning deep features to recognize speech emotion using merged deep CNN." IET Signal Processing 12, no. 6 (2018): 713-721.