A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features

MD ASIM IQBAL[1], MUTHOJU KALYANI[2]
[1]Asst.Professor, ECE Department, Kakatiya University, Warangal, Telangana, India.
[2]M.Tech Student, ECE Department, Kakatiya University, Warangal,Telangana, India.

ABSTRACT:

The problem of popularity and classication of feelings in speech is one of the maximum outstanding research subjects, that has gained reputation, in human-laptop interaction within the final a long time. Having recognized the emotions or emotions in human conversations might have a deep effect on information a human's physical and mental state of affairs. This examine proposes a novel hybrid structure based on acoustic and deep features to growth the classication accuracy in the problem of speech emotion recognition. The proposed approach consists of feature extraction, function choice and classication ranges. At rst, acoustic capabilities consisting of Root Mean Square energy (RMS), Mel-Frequency Cepstral Coefcients (MFCC) and Zero-crossing Rate are obtained from voice facts. Subsequently, spectrogram pics of the authentic sound signals are given as enter to the pre-educated deep community architecture, which is VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet and DenseNet201 and deep features are extracted. Thereafter, a hybrid feature vector is created by using combining acoustic and deep functions. Also, the ReliefF algorithm is used to select greater efficient functions from the hybrid feature vector. Finally, so as for the finishing touch of the classication undertaking, Support vector gadget (SVM) is used. Experiments are made the use of 3 famous datasets used in the literature so one can evaluate the effect of diverse strategies. These datasets are Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin (EMO-DB) and Interactive Emotional Dyadic Motion Capture (IEMOCAP). As a consequence, we reach to seventy-nine.41%, 90.21% and eighty-five.37% accuracy charges for RAVDESS, EMO-DB, and IEMOCAP datasets, respectively. The Final outcomes acquired in experiments, clearly, display that the proposed approach is probably utilized to perform the project of speech emotion popularity efficiently. Moreover, while our method is in comparison with the ones of strategies used in the context, it's miles obvious that our technique outperforms others in terms of classication accuracy charges.

1. INTRODUCTION:

Automatic identification of emotions by facial expressions consists of three steps: face recognition, extraction and classification of features or hand movements, facial features, and voice sound that are used to convey emotions and input. Nonetheless, the latest developments of human user interfaces, which have progressed from traditional mouse and keyboard to automated speech recognition technologies to unique interfaces tailored for individuals with disabilities, do not take full account of these important interactive capabilities, sometimes contributing to less than normal experiences When machines were able to understand such emotional signals, they could provide users precise and effective support in ways that are more in line with the desires and expectations of the individual. From psychological science it is generally agreed that human emotions may be divided into six archetypal feelings: shock, terror, disgust, rage, joy and sadness. Facial expression and voice sound play a critical role in communicating certain emotions.

Emotion interpretation has arisen as an essential field of research that can provide some useful insight to a number of ends. People communicate their feelings through their words and facial gestures, consciously or implicitly. To interpret emotions may be used several different types of knowledge, such as voice, writing, and visual. Speech and facial expression have been the valuable tool for identifying feelings since ancient times, and have revealed numerous facets, including mentality. It is an enormous and difficult job to determine the feelings beneath these statements and facial expressions. Scientists from multiple disciplines are seeking to find an effective way to identify human emotions more effectively from different outlets, like voice and facial expressions, to tackle this issue.

Computer intelligence, natural language modelling systems, etc., have been used to gain greater precision in this responsiveness towards various speeches and vocal-based strategies. Analysis of the feelings may be effective in several specific contexts. One such area is cooperation with the human computers. Computers can make smarter choices and aid consumers with emotion recognition and can also aid render human-robot experiences more realistic. We would explore current emotion recognition methods, emotion modelling, emotion databases, their features, drawbacks, and some potential future directions in this study. We concentrate on evaluating work activities focused on voice and facial recognition to evaluate emotions. We studied different technical sets that were included in current methodologies and technologies. The essential accomplishments in

the sector are completed and potential strategies for improved result are highlighted.

2. LITERATURE SURVEY

Facial emotions are important aspects in human communication that help us to understand the intentions of others. Facial expressions convey Non-verbal Cues which play an important role to maintain interpersonal relations. According to different surveys verbal component (speech) convey one-third of human and Non-Verbal components (Facial emotions, Gestures) convey two-third of human communication. Facial emotion detection became a well attempted research topic now days due to its prospective accomplishments in many domains such as Medical engineering, Vehicles, Robotics and Forensic applications etc. Emotion Recognition will help to understand the inner feelings for people by using their facial expression.

[1] 2017 IEEE 4th InternationalConference on KnowledgeBased Engineering and Innovation (KBEI) https:/ /ieee xplore. ieee.org/document/8324974

[2] Communication 2019 1st on Innovations in Informationand Communication Technology (ICIICT). https://ieeexplore. ieee. org/document/8741491

[3] International Journal of MachineLearning and Computing, Vol.9, No. 1,February2019http://www.ijmlc. org/ vol9/759-L0 179.pdf

[4] D Y Liliana, Published under licence by IOP Publishing Ltd Journal ofPhysics: Conference Series, Volume1193, conference https://iopscience.iop.org/article/10.1088/1742-6596/1193/ 1/ 0 1 2004.

Existing System

In the existing system affective computing is the “computing that relates to, arises from, or influence emotions”, or in the other words, any form of computing that has something to do with emotions. The creation of automatic classifier involves collecting information, extracting the features which are important and finally training the data, so it classify and recognize some patterns. To build a model have to extract emotion of happiness and sadness from facial expression and have to feed the model with pictures of people smiling, tagged with ”happiness”, and with pictures of people frowning, tagged with “sadness”. After that, when it receives a picture of a person smiling or frowning, it identifies the shown emotion as “happiness” or “sadness”. Emotion detection using speech, gathering emotional information from the user of a system is their voice. Any emotion from the speaker’s speech is represented by the large number of parameters which is contained in the speech and changes in these

parameters will result in corresponding changes in emotions which is quite difficult.

Disadvantages of existing system

• Creation of model in real life is difficult.

• Voice recognition software won't always put your words on the screen completely accurately.

• Programs cannot understand the context of language the way that humans can, leading to errors that are often due to misinterpretation.

3. METHODOLOGY

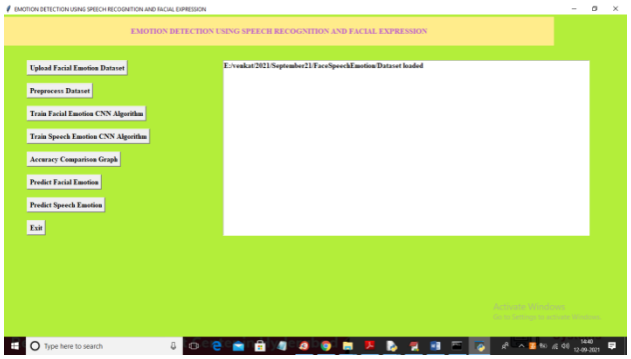
To overcome the existing drawbacks, comparing the traditional machine learning approaches, deep learning based methods have shown better performance in terms of accuracy and speed of processing in image recognition. We have used a modified Convolutional Neural Network (CNN). CNN is mostly used in image and face recognition. CNN is a kind of artificial neural networks that employs convolutional methodology to extract features from the input data to increase the number of features from live video streaming. That captures each frame and test them and is trained by CNN model and later classified into different emotions. With computational power of Graphical Processing Units (GPU's), CNN has achieved remarkable cutting edge results in image recognition.

Emotion and Features:

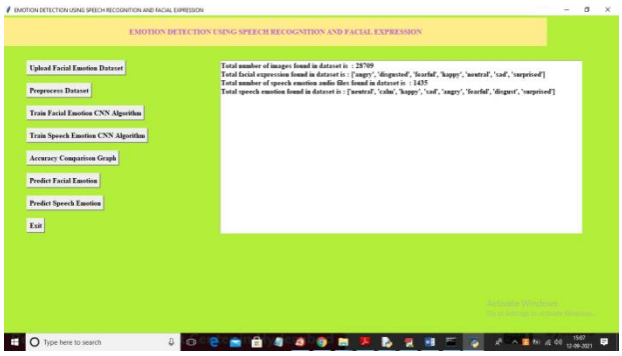
Psychological studies have shown that changes in human emotions reflect through prosodic parameters of speech. Generally, acoustic features associated with the emotions including pitch, duration time, energy, formant, and average, maximum, minimum, intermediate values, ranges, the first derivative, the second derivative and change rates derived from them [7]. After repeated experiments, this paper eventually selected the following prosodic features: phonation time, speech rates, basic frequency averages, basic frequency ranges, basic frequency change rates , Amplitude averages, Amplitude change ranges, formant change averages, formant change ranges, and formant change rates.

Face features generally include three kinds of: Geometric features, physical features, mixing features. The physical features refer to the features using the whole face image pixels, reflect the underlying information of face images, and focus on extracting the subtle changes of local features [8]. However, the number of feature point? extracted is too many that resuling to the higher dimension and the complex calculations. Mixing features combine the geometric features with physical features. The calculation of it is also complex, and the initial point is difficult to obtain [9]. The recognition effect of the geometric features requires a higher accuracy of the Datum point extracted. The recognition effect

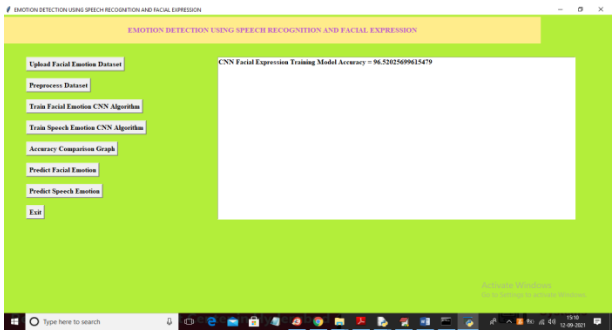
of requiring a higher accuracy of the Datum point extracted. Meanwhile extracting the geometric features ignores the other information of faces (such as skin texture changes etc.) But it can describe the macro structural changes of the face, and the easy way to extract and the lower dimension making it quite comply with the requirements of our emotional system.



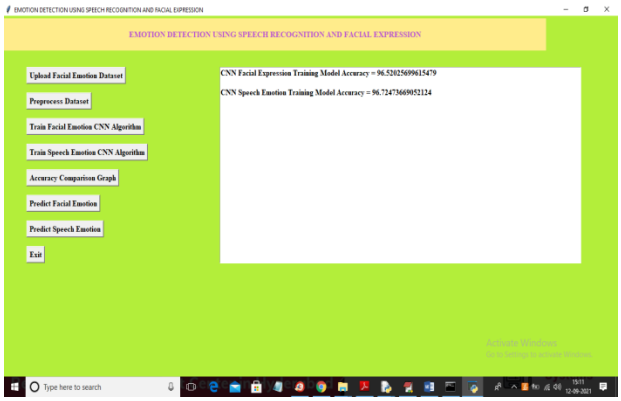
In above screen dataset loaded and now click on 'Preprocess Dataset' button to read all images and then resize them to equal size and then extract MFCC features from dataset and then build trained model.



In above screen both datasets are processed and we can see total number of images and audio files available in both datasets and now dataset is ready and now click on 'Train Facial Emotion CNN Algorithm' button to train Facial dataset with CNN and to get below screen

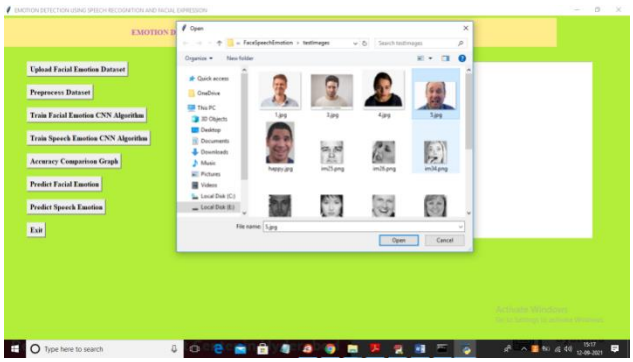


In above screen training CNN with Facial images got 96.52% accuracy and now click on 'Train Speech Emotion CNN Algorithm' button to train CNN with audio features and to get below output

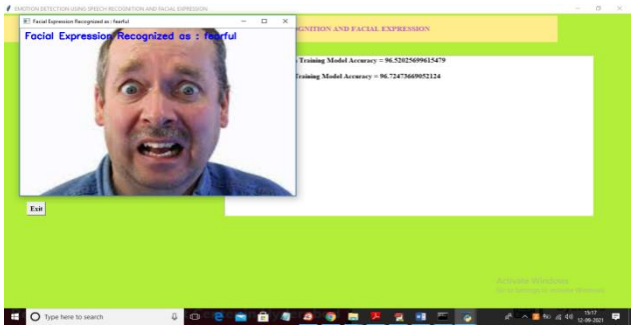


In above screen with CNN speech Emotion we got 96.72% accuracy. Now click on 'Accuracy Comparison Graph' button to get below graph

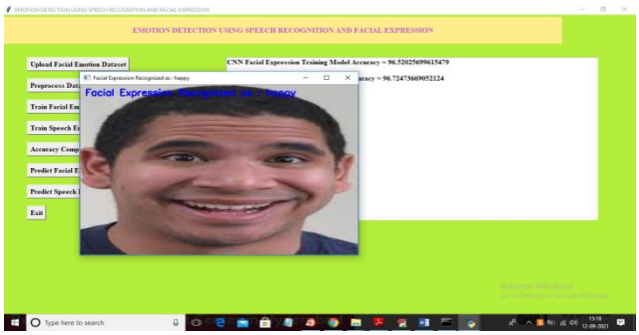
In above graph x-axis represents EPOCH and y-axis represents accuracy and loss values and we can see both algorithms accuracy reached to 1 and both algorithms loss values reached to 0. In above graph green line represents face emotion accuracy and blue line represents speech accuracy. Now click on "Predict Facial Emotion" button to upload face image and will get below result



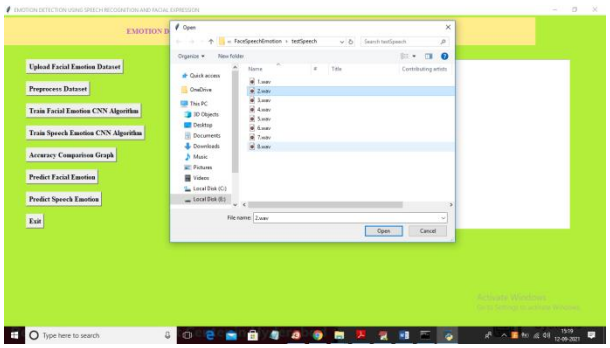
In above screen selecting and uploading '5.jpg' image and then click on 'Open' button to get below result



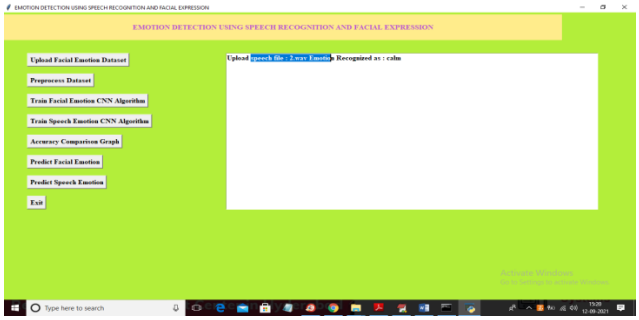
In above screen facial emotion or expression predicted as 'Fearful' and now test other image



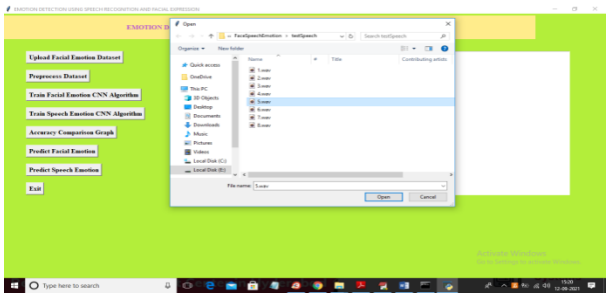
In above screen facial expression predicted as 'happy' and similarly you can upload other images and test. Now click on 'Predict Speech Emotion' button to upload audio file and get below result



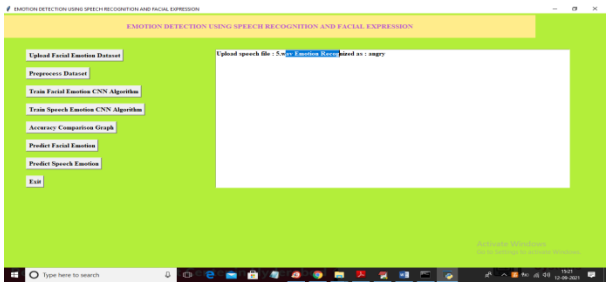
In above screen selecting and uploading '2.wav' file and below is the result



In above screen uploaded audio file emotion predicted as 'calm' and now test other file



In above screen uploading '5.wav' file and below is the prediction result



In above screen uploaded file emotion predicted as 'angry' and similarly you can upload other files and test.

CONCLUSION

In the experiment, we analyze and compare time, amplitude energy, basic frequency and formant feature parameters under different emotional states, and find out the distribution laws of different emotional signal features. On this basis, we classify five emotional states of calm, sadness, happiness, surprise, and anger. The recognition results show that on these basic Prosodic information we can initially recognize basic emotional catagories, and apply it into the emotion recognition system, which limits the amount of storage and computation and doesn't have strict recognition accuracy. Meanwhile, the prosodic features, integrating with facial expression information, recognizes emotional categories with Bimodal, reaching a higher recognition rate.

Although the emotional recognition performance combinating with facial expressions has improved, The recognition rate doesn't improve significantly. This is mainly because in the terms of obtaining the emotional information, there is a similar correlation between the adjacent video frames, due to the continuity of the facial expression changes. But we didn't take this correlation into consideration when catching the instant face image to analyze separately. On the other hand, when the facial expression changes, the shape and the location of the organs on the face, will change accordingly. In this paper,although the image analysis method based on Gaussian mixture algorithm has a higher recognition rate for the face contour, it lacks of detailed characterization of changes in the eyes, nose, mouth and other facial organs. Based on the above two reasons, in order to truly improve the system performance, we need to build a correction model associated with the expressions containing a variety of rules, and modified the image recognition results using the model. In addition, in the term of real-time applications, besides enhancing the robustness of the system and improving the accuracy, the efficiency of the

recognition algorithm is also a key factor. The strategies such as codebook pruning, data compression can also improve the recognition rate effectively

Multi-modal recognition systems intergrating with images, voice and other emotional information is the inevitable trend of future human-computer interaction development. Although there are still many insurmountable technical problems, with the continuous progress of science and unremitting efforts of the researchers, the real-time systems of multi-modal speech recognition will have more potential development.

REFERANCES

[1] Lingxiao He, Haiqing Li, Qi Zhang, Zhenan Sun Member, IEEE, "Dynamic Feature Matching for Partial Face Recognition", This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2870946, IEEE Transactions on Image Processing

[2] G. E. Hinton, I. Sutskever, A. Krizhevsky, "Deep convolutional neural networks With imagenet classification ".

[3] T.Ahonen ; A.Hadid ; M.Pietikainen,"Face Description with Local Binary P atterns: Application to Face Recognition",IEEE Transactions on Pattern Analysis and Machine Intelligence Year: 2006 , Volume: 28 , Issue: 12

[4] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. "Deep face recognition". In BMVC, pages 1–12, 2015.

[5] D. Kalenichenko, J. Philbin, F. Schroff, "Face recognition and clustering using a unified embedding ."

[6] Y. Sun, X. Wang, and X. Tang, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),"Deep learning face representation from predicting 10000 classes", pages 1891–1898, 2014.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, pages 1701–1708, Deep face: Closing the gap to human-level performance in face verification", 2014.

[8] Dayong Wang, Member, IEEE, Charles Otto, Student Member, "Face Search at Scale", IEEE "Transactions on Pattern Analysis and Machine Intelligence Year: 2017"

[9] IsmahaneCheheb, NoorAl-Maadeed , SomayaAlMadeed , Ahmed Bouridane ,Richard Jiang, "Random sampling for patch based face recognition",2017 5th International Workshop on Biometrics and Forensics (IWBF)Year: 2017

[10] RaviSubban , SavithaSoundararajan,"Human face recognition using facial feature detection techniques", 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) Year: 2015

[11] Lingxiao He , Haiqing Li , Qi Zhang , Zhenan Sun , Zhaofeng He,"Multiscale representation for partial face recognition under near infrare d illumination"2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)Year: 2016

[12] ShengcaiLiao ,AnilK. Jain , Stan Z. Li,"Partial Face Recognition: AlignmentFree Approach",Transactions on Pattern Analysis and Machine Intelligence Year: 2013 , Volume: 35 , Issue: 5

[13] Evan Shelhamer; JonathanLong ; TrevorDarrell"Fully Convolutional Networks for Semantic Segmentation", IEEE Transactions on Pattern Analysis and Machine IntelligenceYear: 2017 , Volume: 39 , Issue: 4

[14] RenliangWeng ,JiwenLu , Yap-PengTan, "Robust Point Set Matching for Partial Face Recognition", IEEE Transactions on Image ProcessingYear: 2016 , Volume: 25 , Issue: 3

[15] R. Brunelli and T. Poggio, IEEE Transactions on Pattern Analysis and Machine Intelligence, "Face recognition: features versus templates," 1993.