International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.163

Speech Emotion Recognition Using Deep Learning Algorithm

Ipsita Roy Barman
MasterofTechnology
DeptofDataScienceandBusinessSystem
SRMInstituteofScience&Technology
Chennai, India

Dr.AShanthini
DeptofDataScienceandBusinessSystem
SRM Institute of Science & Technology
Chennai, India

***Abstract*—AI is the emerging technology and is being extensively used now a days in every field and in every organization to gain more insights and hidden patterns in data. Data is the new oil for every organization to run and make huge revenues if one is able to find out the story in the past data and then can recommend based on it in the future. AI not only works effectively in data science related fields but is now taking grip for every work fields as well. AI devices become more interesting when they start to act more like human rather than robotic as the customer interactions increase and so does the revenue of the organization. So, emotion recognition of the users is a very important partin AI development where the device is trained to recognize the correct emotion of the customer behind his / her speech and conversation with the device. We know that random forest uses a bunch of decision trees and tries to correct the errors foundinthepreviousstepstorectifythemandthen increase the accuracy. In this article, we use CNN basedmodelwhichisaverypopulardeeplearning neural network-based algorithm to classify every emotion of an individual and then used audios to test the model. In this project, CNN model is used and 79.6% accuracy is achieved.**

***Keywords:* Speech emotion, deep learning, CNN, Artificial Intelligence, Decision trees, Random Forest.**

I. Introduction

Speechandwordsarethemostnecessarytool we use to interact with each other. Now a daysAIhasbecomeanintegralpartofourlife speciallytheAIbasedtechnologies. Welove to interact with them the whole day and give themrequestsortalksomerandomthings.As a result, we as a customer expect some interestingrepliesfromthemandthattoolike a human being with emotions. In this fast- changing world and generation we are left with a few friends and we have no time to meet or interact with them. So, in this new world,AIisournewfriend,philosopher,and guide too in many cases. Hence our inputs andcommandsareconsideredasdataforthe above-mentioned problem statement and with the help of CNN we make the audio tuning more perfect and the AI more lifelike. The model is trained in such a way to recognise emotions behind every input of the customers.

1.1. *Scope*

Data is the future of any organisation and thus AI is the emerging technology which uses data as the input and in returngivesamazingexperiencetoitscustomers. Themajor risk faced in the usage of AI is constant fall of customers interestinusingit.Thisissuecanbesolvedeasilybymaking the output responses of AI more interesting and natural sounding.For thisthingtohappen,weneed torecognise the emotionbehindthecustomersandthenteachtheAItodothe same thing. Once it will be able to recognise the emotion behindthespeech,itwillresponseinamorelucidandhuman like. The neural networks in architecture like CNN works similar to our brain neural structures with the nodes to transfer data from one layer to another. Similarly in neural network,dataispassedfrominputtooutputlayersthrougha series of hidden layers in between. Hence, this method is proposed in this project so that the AI can be transformed successfullyintoacompletehumanlikestructure.Ourmodel will not onlyrecognisetheemotions behind everyspeech of our customers but can also cheer the customers when they willbesadordepressed.Oncethismodelisbuiltsuccessfully with great precision and accuracy and minimum errors, we can collect data of our customers to classify them into depressed and non-depressed individuals based on the durationoftheirpastdataofbeinginnegativeemotionssuch as sadness, anger, fearful, disgust, etc. This entire process will not be possible without the technologies and machine learning algorithms cause data is involved. Data will in fact act as the raw material for this project, without which progress of anykind in this work is just futile. In futurethis model can be applied on any AI devices or AI applications and collecting customers data from the devices will be very important and can be used to recommend various things for their future needs.

1.2 *Methods*

1.2.1 DeepLearning

Thisbranchofmachinelearningisveryuseful asitinvolves the use of neural networks along with the multiple layers and can process complex data sets. It even is capable for developing artificial intelligence systems that can learn on their own, keep on improving themselves on their own without explicitlybeing programmed by us for a particular task. The algorithms can recognize patterns and making predictions based on large amountofdataandthisfeatureindeed isusedin therecognition of emotions behind the speech of our end users.

1.2.2 CNN

It is a type of neural network that is commonly used in image and video analysis tasks and as it has various convolutional layers so it can learn various spatial hierarchies of features and thisin turn canscan theinput datawith aset offilterstoextract relevant features. They typically consist of convolution layers, poolinglayers,andfullyconnectedlayers.Theyareresponsibl e for detectinglowlevel featuresandpoolinglayers downsample the feature maps toreduce the

DOI: rp-9788770040723.163

dimensionalityof the data. Fully connectedlayers combinethe extracted featurestomakea final prediction. As they are responsible for processing structured data, this algorithm is useful in extracting the correct emotion out of the audios we used in this project.

II. LITERATURESTUDY

Many studies on speech emotion recognition has been conductedandvariousnewalgorithmsandmethodshasbeen proposed by authors. This is a rapidly growing field of research which attracted significant attention from researchers in recent years. Let's look at some of the important significant reviews:

S.Bhattacharyya and S. Poria [1] in 2017 provides a comprehensive review of various techniques and methods anditdiscusses the challenges associatedwiththetasksuch as the standardizeddatabase for emotions is lackingandthe need for robust feature extraction techniques.

VBalakrishnanandS.Sivakumar[2]in2019proposesanew method for emotion recognition using speech features such as pitch, energy, and formats. The proposed method was tested on the Berlin Emotional Speech Database and achieves an accuracy of 88.3%.

P.Rao and P.S.S. Avadhani [3] in 2020 proposed recent advances in deep learning based approaches for SER and discussed the use of RNN for feature extraction and classification. The paper also covers the use of transfer learning.

A. Gunavati and R. Bhavani [4] in 2021 presented a comprehensive survey of deep learning techniques and covers recent developments in multimodel emotion recognitionwhichinvolvescombiningspeechwithother modalitiessuchasfacialexpressionsandphysiological signals.

R. Mitra [5] in 2019 presented a paper which contains a comprehensive survey of various techniques and methods used for SER., including traditional machine learning approaches, deep learning techniques and hybrid models.

T.A.Rahman[6]in2020fine-tunedthepretrainedmodelon theRyersonAudioVisualdatabaseofemotionalspeechand song dataset using transfer learningtechnique andachieved 70% accuracy.

M.A.H.Akanda[7]in2020investigatedtheeffectofspeech enhancement techniques on SER using deep neural networks.Theauthorcomparedtheperformanceofdifferent enhancement methods such as spectral subtraction, and MMSE-based noisereduction, on the Emo-DB dataset. The results show that speech enhancement can significantly improve the accuracy of emotion recognition.

W. Wang [8] in 2021 proposed a multi task learning approach for SER that integrates both acoustic and lexical features. Theauthorsuseadeepneural networkwithshared and task-specific layers to jointly learn the feature representations for emotion recognition and sentiment analysis.

G.F. Adewumi [9] in 2021 presented a review of various features, classification techniques, and datasets used for SER. It discusses the challenges associated with SER such as the variability of emotions across speakers and cultures.

M.R. Islam [10] in 2020 compared the performance of various deep learning-based approaches for SER, including LSTM networks and hybrid models. The authors evaluated themodelsontheIEMOCAPdatasetandshowedthathybrid model perform the best.

III. PROPOSEDMETHODOLOGY

3.1 SystemProposed

Deepneuralnetworks(DNNs)areatypeofmachinelearning model that has gained significant attention in recent years due to their ability to learn complex patterns and representations from large datasets. DNNs have shown promising results in various applications, including computer vision and speech emotion recognition.

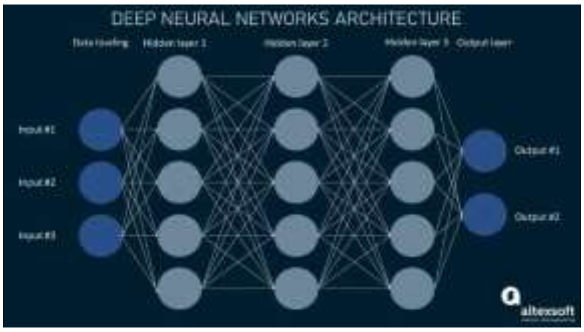

Fig1.DeepNeuralNetwork

In otherwords, DNNcanberepresentedasa functionf(x;a) wherexis the input, a represents the model parameters, and f(x;a) represents the output of the network for a given input x.

3.2 Dataset-Taken

Thedatasetcontains1440files,7356recordings,24professionalactors who acted out different emotional states including, calm, happy, sad, angry, fearful, surprise and disgust.

3.3 DatasetPreprocessing

Dataset pre-processing is a crucial step in machine learning anddataanalysisthatinvolvespreparingrawdatatobeused as input to a model. The goal of pre-processing is to transform the raw data into a format that can be easily analysed and interpreted by a machine learning algorithm. Data pre-processing steps include data cleaning, data normalization, data transformation, data encoding, data splitting and feature selection.

3.3.1 AugmentationOfData

DOI: rp-9788770040723.163

It is a technique used to artificially increase the size of a datasetbycreatingnew,butsimilar,versionsofexistingdata. Thegoalofdataaugmentationistoimprovetheperformance of machine learning models by providing more diverse examples for the model to learn from. This is particularly usedwhenworkingwithsmalldatasetswhereoverfittingcan be a concern. There are various methods of data augmentationandthespecifictechniquesuseddependonthe type of data being augmented. The purpose of data augmentation is to create new examples that are like the original data but not identical. The augmented data should represent the natural variations in the data that the model is likely to encounter in real-world scenarios. Data augmentation is an important tool in machine learning and can significantly improve the performance of models, especially when working with small datasets. The idea behind data augmentation is that by creating additional training samples that are similar but not identical to the original data, a model can better generalize to new, unseen data. This can help to reduce overfitting and improve the overallperformanceofthemodel.Dataaugmentationis commonly used in computer vision and natural language processing tasks, but it can also be applied to other types of data such as audio or sensor data. The specific techniques used for data augmentation depend on the type of data and the specific task at hand, and it is often necessary to experiment with different augmentation methods to find the best approach for a particular problem statement. Data augmentation is typically done by applying a set of transformation rules to the existing data samples, such as rotating, flipping, or zooming in on images or changing the pitchorspeedofaudiosamples.Sothisdataaugmentationis an irreplaceable step in this project.

3.4 Deep-Learning

Deep learning for speech emotion recognition refers to the use of deep neural networks to classify and analyze emotional states in speech signals. The goal of SER is to automatically detect the emotional content of speech, such as happiness, sadness, anger, or fear. It typically involves multiple layers of artificial neurons that aretrainedonlargeamountsoflabelleddatatoidentifypatterns and relationships between different features of speech signals, such as pitch, duration, and spectral characteristics. These models can be trained using various architectures such as Convolutional Neural Networks, RNN, or hybrid model that combines both RNN and CNN. The training process for deep learning models involves feedingthenetworkwithlargeamounts oflabelledspeechdataand adjustingtheweightsofthenetworksparametersthrough aprocess called backpropagation in order to minimize the error between the predictedemotionallabelsandthetrue labels.Theresultingtrained model can be used to classifythe emotional content of newspeech signals. Deep learning models for SER have shown promising results in recent years and have been used in a variety of applications. However there are still many challenges to be addressed in this field such as dealing with variability in speech signals, addressing class imbalance in emotional labels, and improving the interpretability of the models.

3.5 ConvolutionalNeuralNetwork(CNN)

TheCNNfor thespeech emotionrecognitiontypicallyinvolves severallayersthataredesignedtoextractrelevantfeaturesfrom speech signals and then classify the emotional content.

1. Inputlayer:thislayerreceivestherawspeech signalas an input. This layer can be preprocessed with techniquessuchasMel-frequencycepstralcoefficients or filter bank energies to extract relevant features.

2. Convolutionallayer:thislayerappliesasetoffiltersto the input signal which extracts local features such as pitch and spectral information.

3. ReLU activation layer: this layer applies a non-linear transformationtotheoutputoftheconvolutionallayer, introducing non-linearity into the model.

4. Pooling layer: this layer performs a downsampling operation on the output of the ReLU layer, reducing spatialdimensionsofthefeaturemapandincreasingthe model’s robustness to small variations in the input signal.

5. Dropout layer: it randomly drops out a fraction of the activations in the previous layer during training, preventingoverfittingandimprovingthegeneralization of the model.

6. Fully connected layer: This layer takes the flattened output of the previous layer and applies a linear transformation to it, producing a set of scores that represent the probability of each emotion class.

7. Softmaxactivationlayer:appliesasoftmaxfunctionto the output of the fully connected layer, producing a probability distribution over the possible emotion classes.

8. Outputlayer:theoutputlayeroftheCNNproducesthe final predicted emotion class based on the probability distribution from the softmax layer.

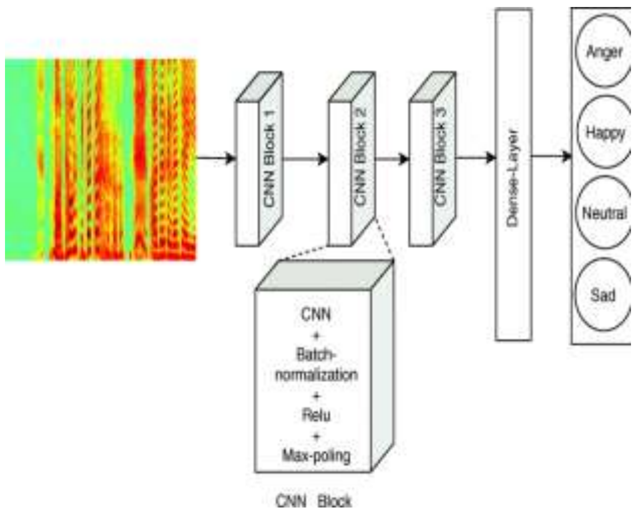

Fig.2.Architecture

IV. MODULES

4.1 CreationOfModels

Creatingmodelsforspeechemotionrecognition using CNN typically involves the following steps:

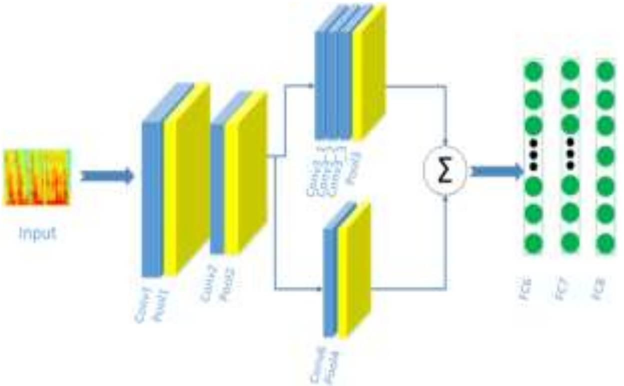

Fig3.Model Diagram

Ourapproachconsistsoffivesteps:

Step1–Datapreprocessing:

This involves converting the raw speech signal into a format that can be used by the network, such as Mel-frequency cepstralcoefficients(MFCC) or filter bank energies.

Step2-Dataaugmentation:

ThisisdonetoimprovetheperformanceoftheCNNand is often helpful to augment the training data by creating additional synthetic samples.

Step3-Modelarchitecturedesign:

The next step is to design the architecture of the CNN. This is typically done to involve selecting the number and size of convolutional and pooling layers, choosing activation functions,anddecidingonthenumberoffullyconnectedlayers.

Step4-Trainingthemodel:

Oncethemodelarchitectureis defined,thenextstepisto train the model on the training data. During training, the weights of the network are adjusted to minimize the error between the predicted emotional labels and the true labels.

Step5-Modelevaluation:

After training, the performance of the model is evaluated using a separate test set. This allows the accuracy, precision,recall,and F1scoreofthemodeltobecalculatedand compared with other models.

Step6-Fine-tuning:

If the performance of the model is not satisfactory, it may be necessary to fine-tune the model by adjusting the architectureorhyperparameters.Thiscanbedonebyevaluating theperformanceofthemodelonthevalidationsetandadjusting accordingly.

Step7- Deployment:

Once the model has been trained and evaluated, it can be deployed in a production environment to perform real-time SER tasks.

To evaluate the optimal efficiency and robustness of the algorithm, metrics such as Precision and Recall rates are evaluated and computed based on the recognition rate.That the proposed system produces the highest recall rate for all types of parameters like speech andthen findingthe correct emotionbehindthespeech.. Theaverage ofallmeasuresfor the proposed system.

Overall creating models for SER using CNN s involves several steps from data preprocessing to modelarchitecture design, training, evaluation, and deployment. Each step requires careful consideration and experimentation to achieve optimal performance.

4.2 SpeechEmotionRecognition

Thedeployedmodeltakesinnewspeechsignalsandpredictsthe emotional content of the speech in real-time. Overall, SER involvescollectingandpreprocessingdata,augmentingthe data,training and evaluating a deep learning model, fine-tuning the model, and deploying the model for real-time SER tasks.

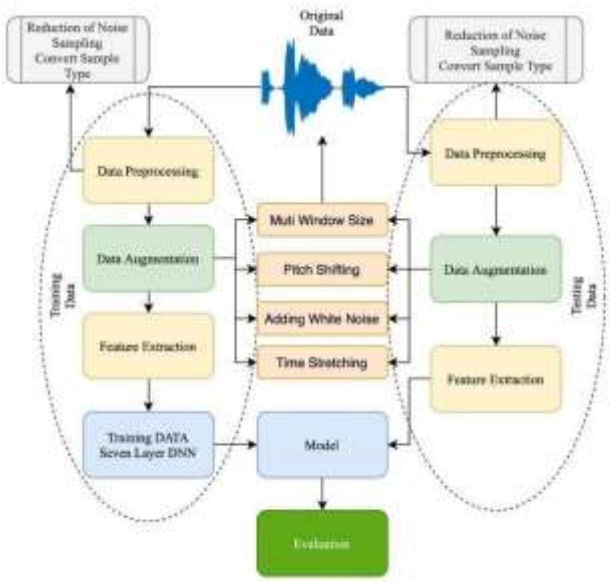

Fig:4Activity Diagram

4.2.1 Performance o f a Speech Emotion Recognition:

Theperformancecanbeevaluatedusingvariousmetricssuch as accuracy, precision, recall, and F1 score. Accuracy measures the percentage of correctly predicted emotional labelsoutofallthelabelspredictedbythemodel.Precision measures the percentage of correctly predicted positive emotionallabelsoutofallthepositivelabelspredictedbythe model.Recallmeasuresthepercentageofcorrectlypredicted positive emotional labels out of all the positive emotional labels in the dataset. F1 Score is the harmonic mean of precisionandrecall.TheperformanceoftheCNNmodelcan befurtherimprovedbyfinetuningthemodelarchitectureand hyperparameters augmenting the training data, and using transferlearningtechniquestoleveragepre-trainedmodelsin larger datasets.

DOI: rp-9788770040723.163

TABLE1:PERFORMANCEOFSERSYSTEMUSINGCNN

Actual Class	Predicted Class: Anger	Sad	Neutral	Happy
Anger	43.3	12.6	33.7	10.4
Sad	9.6	78.3	0	12.1
Neutral	3.6	0.3	93.3	2.8
Happy	25.9	0	27	47.1

Architecture

Additionally, the performance of the system can be evaluated on different test sets and compared to other state-of-the-art models to determine its effectiveness in recognizing emotions from speech signals.

Considering a happy track from the dataset with plt.figure explaining the figsize to be (15,5).The model was able to recognisethecorrectemotionbehindtheaudioandwaseven able to figure out the gender of the speech.

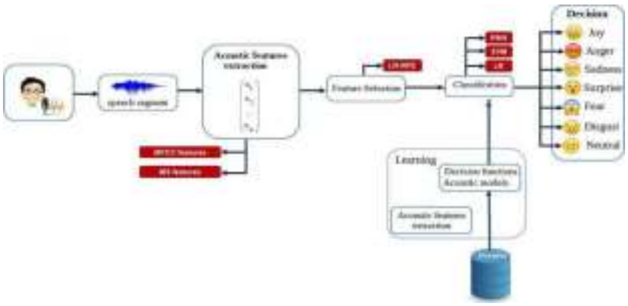

Fig5:Sequence Flowchart

GENDER DETECTION: it involves using a machine learning model to predict the gender of the speaker based on the emotional content of their speech. This is achieved by combining deep learning models like CNNandRNN.TheCNNisusedtoextracthigh-levelfeaturesfromthe speech signal, while the RNN is used to model the temporal dynamics of the speech signal. Once the deep learningmodel has been trained on adatasetoflabelledspeechsamples,itcanbeusedtopredictthegender ofnewspeechsamples.Themodelwasthenfedapre-processedspeech signal, extracted the relevant features using the CNN, and then these featureswerepassedthroughRNNtomodelthetemporaldynamics.The output of the RNN is then fed into a fullyconnected layer that predicts the gender of the speaker.

4.2.2 DecisionMaking

Basedontheinputdataintothemodelitwasdecidedwhetherthepperson was a male or female and the emotion behind the speech of that individual. This step involves making a prediction for the emotional state of the speaker based on the features extracted from the speech signalbytheCNN.Aftertheinputspeechsignalhasbeenpreprocessed and transformed into a format that can be used by the CNN, the CNN extracts high-level features from the speech signal using a series of convolutionallayers.ThefeatureslearnedbytheCNNarethenfedinto afullyconnectedlayerthatmakesthefinalpredictionfortheemotional state of the speaker. During the decision-making step, the CNN takes the preprocessed speech signal as input, extract the relevant features using the convolutional layers, and passes these features through the fullyconnectedlayertomakeapredictionfortheemotionalstateofthe speaker. The predicted emotional label can then be used for further analysis ortocontrol other systemsbasedonthe emotional state ofthe speaker.

Overall, the decision-makingstep-in SERusingCNNinvolves usinga fullyconnectedlayertomakeapredictionfortheemotionalstateofthe speaker based on the features extracted from the speech signal by the CNN.

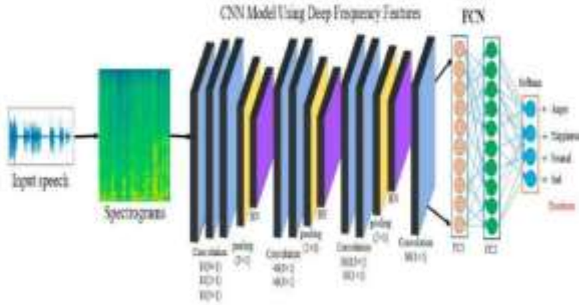

Fig 6:Decision Making Steps In The Speech Emotion Recognition

The speech component of the dataset includes 1440 recordings of 60 sentences spokenin 8 different emotional expressions whilethe song component includes 1012 recordings of 52 songs sung in 7 different emotional expressions. Each recording is labelled with metadata indicating the actor, gender, expression, and type of recording. The dataset is designed to be used for research and developmentinareassuchasemotionrecognition,speech processingandaudioanalysis.Wecarriedoutalltheseprocessesin the project.

4.2.3 FineTuningThe Model

The first step was to load the pre-trained CNN model and remove the last layer, which is typically the output layer that is specific to the original task the CNN was trained on, such as image classification. The remaining layers are frozen, meaning their weights are not updated during training.

Thenextstepwastoaddanewoutputlayer,whichisspecifictothe sentiment analysis task, with the appropriate number of output nodes for thenumber of sentiment classes. The weights ofthenew output layer are randomly initialized.

Finally, the entire network is fine-tuned by continuing the training process with a smaller learning rate. This allows the network to adapt tothedomain-specificsentiment dataset and furtherimprove its accuracy.

Fine-tuning a pre-trained CNN can greatly reduce the time and computationalresourcesrequiredtotrainanewmodelfromscratch while still achieving high accuracy in the domain-specific task of the speech emotion recognition.

International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.163

V. Results

Theobtainedresultssuggestedthatthedatasetisavaluable resource for training and evaluating models for speech emotion recognition, and that there are several effective approaches for achieving high accuracy on this task.

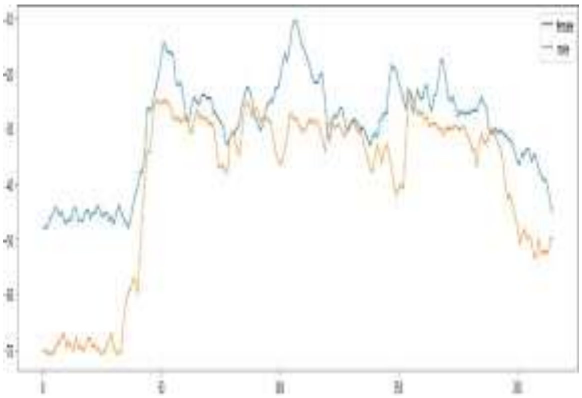

Fig7:AccuracyGraph

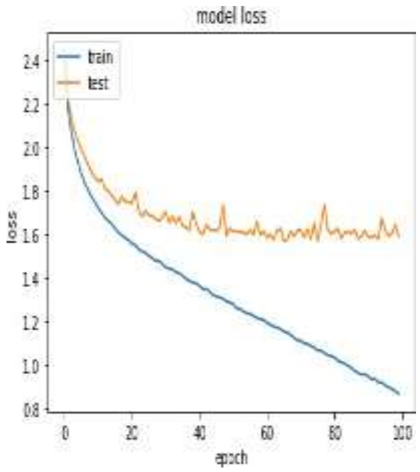

Fig 8:ModelLossvs no.Ofepochs

	actualvalues	predictedvalues
170	male_sad	male_sad
171	female_disgust	female_disgust
172	male_angry	male_happy
173	female_disgust	female_disgust
174	male_angry	male_angry
175	female_fear	female_happy
176	male_disgust	male_surprise
177	female_fear	female_happy
178	female_happy	female_happy
179	female_neutral	female_neutral

Fig 9: Actual Vs PredictedValues

Fig.8 and Fig.9 shows that 4 out of 10 random state of thedatasetshowsincorrectlabelswherefinetuningof model come into picture.

TABLE2:TrainingResultsForMaleAnd FemaleSamplesUsingRBFNetwork

Sample name (Input - Male)	Pitch Observed	Gender Classified Correctly	Emotion Frame Count	Emotion Classified Correctly
Angry	108.0	Yes	255.6	Yes
Angry	188.0	Yes	267.8	Yes
Angry	67.0	Yes	273.9	Yes
Angry	171.0	Yes	250.3	Yes
Angry	72.0	Yes	265.1	Yes
Sad	35.0	Yes	35.9	Yes
Sad	347.0	No	46.8	Yes
Sad	89.0	Yes	78.9	Yes
Sad	171.0	Yes	198.4	No
Sad	188.0	Yes	76.9	Yes
Neutral	209.0	Yes	102.4	Yes
Neutral	134.0	Yes	143.3	Yes
Neutral	190.0	Yes	127.8	Yes
Neutral	455.0	No	109.2	Yes
Neutral	98.0	Yes	145.3	Yes
Happy	189.0	Yes	167.8	Yes
Happy	112.0	Yes	153.4	Yes
Happy	69.0	Yes	155.9	Yes
Happy	41.0	Yes	89.0	Yes
Happy	45.0	Yes	176.3	Yes

Sample name (Input - Female)	Pitch Observed	Gender Classified Correctly	Emotion Frame Count	Emotion Classified Correctly
Angry	388.0	Yes	298.3	Yes
Angry	331.0	Yes	312.5	Yes
Angry	121.0	No	303.2	Yes
Angry	345.0	Yes	278.9	Yes
Angry	72.0	No	289.4	Yes
Sad	335.0	Yes	35.9	Yes
Sad	347.0	Yes	79.0	Yes
Sad	389.0	Yes	59.6	Yes
Sad	341.0	Yes	298.1	No
Sad	338.0	Yes	37.8	Yes
Neutral	393.0	Yes	105.6	Yes
Neutral	334.0	Yes	139.8	Yes
Neutral	397.0	Yes	125.6	Yes
Neutral	355.0	Yes	133.5	Yes
Neutral	348.0	Yes	126.7	Yes
Happy	459.0	Yes	187.2	Yes
Happy	112.0	No	176.3	Yes
Happy	469.0	Yes	199.0	Yes
Happy	451.0	Yes	155.8	Yes
Happy	435.0	Yes	167.8	Yes

Advantages

The various advantages of the implemented method or system are:

1. Improving Human- compute interaction.
2. Improving mental health.
3. Improving customer service.
4. Enhancing security.
5. Advancing scientific research.

Conclusion

Weproposedanewmethodtoimprovementalhealthofan y person by detecting changes in emotional state over time, which could be helpful in identifying individuals

DOI: rp-9788770040723.163

who may be at risk for mental health issues such as depression or anxiety. Wewill becollectingall thepast data of customers overagoodperiodoftime,analysethedataandwillbeable to figure out who are at risk of calling themselves as depressed or non-depressed person. This is a rapidly developingfield with a wide range of potential applications in various industries and fields. As technology continues to advance and more data becomes available, we can expect that SER models will continue to improve, providing more accurate and reliable detection of emotions in speech.

FUTUREWORK

Thereareseveraldirectionsthatfutureresearchin speech emotion recognition could take:

1. Incorporating contextual information.
2. Improving model robustness.
3. Handling multilingual speech.
4. Recognizing more complex emotions.
5. Integrating with other technologies.

REFERENCES

[1] M.ElAyadi,M.S.Kamel,andF.Karray,“Surveyonspeech emotionrecognition:Features,classificationschemes,anddatabases,” PatternRecognit.,vol. 44, no. 3, pp. 572–587, Mar.2011.https://doi.org/10.1016/j.patcog.2010.09.020Crossref

[2] R.BanseandK.R.Scherer,“Acousticprofilesinvocalemotionexpression.,” Pers. Soc. Psychol, vol. 70, no. 3, pp. 572–587,1996.https://doi.org/10.1037/0022-3514.70.3.614GoogleScholarCrossref

[3] V.HozjanandZ.Kačič,“Context-IndependentMultilingualEmotionRecognitionfromSpeechSignals,” Int.J.SpeechTechnol.,vol.6,no.3,pp.311–320,2003.https://doi.org/10.1023/A:1023426522496GoogleScholar Crossref

[4] F. Burkhardt, A.Paeschke, M.Rolfes, W. F.Sendlmeier,and B. Weiss, “AdatabaseofGerman emotional speech.,”inInterspeech,2005,vol.5,pp.1517–1520.GoogleScholar

[5] A.Schuller,B.,Steid,S.l,andBatliner,“Theinterspeech2009emotionchallengee,” Interspeech, pp. 312–315, 2009.GoogleScholar

[6] M.M.H.El Ayadi,M.S.Kamel,andF.Karray,“SpeechEmotionRecognitionusing GaussianMixtureVectorAutoregressiveModels,”in2007IEEE InternationalConferenceon Acoustics, Speechand Signal Processing -ICASSP ’07, 2007, vol. 4, pp. IV–957–IV–960.GoogleScholarCrossref

[7] W.Dai,D.Han,Y.Dai,andD. Xu,“EmotionRecognitionand AffectiveComputing on Vocal Social Media,” Inf. Manag., Feb. 2015.GoogleScholar

[8] B. S. Atal,“Effectivenessofliner prediction characteristicsofthespeechwaveforautomaticspeakerspeechwaveforautomaticspeakeridentificationand verification,” Acoust. Soc. Am., vol. 55, no. 6, pp. 1304–1312,2005.https://doi.org/10.1121/1.1914702GoogleScholarCrossref

[9] S.Wu,T.H.Falk,andW.-Y.Chan,“Automatic speech emotionrecognitionusing modulation spectral features,” SpeechCommun., vol.53, no.5, pp. 768–785, May 2011.https://doi.org/10.1016/j.specom.2010.08.013GoogleScholar Crossref

[10] H.Cao,R. Verma,andA.Nenkova,“Speaker-sensitive emotionrecognitionviaranking:Studiesonactedandspontaneousspeech,”Comput.Speech Lang., vol. 28, no. 1, pp. 186–202, Jan.2015.https://doi.org/10.1016/j.csl.2014.01.003Google ScholarCrossref PubMed

[11] C.C.Lee,E.Mower,C.Busso,S.Lee,andS.Narayanan,“Emotionrecognition using a hierarchical binary decision tree approach,” SpeechCommun., vol. 53, no. 9–10, pp. 1162–1171, Nov.2011.https://doi.org/10.1016/j.specom.2011.06.004GoogleScholarCrossref

[12] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognitionusinghiddenMarkovmodels,” SpeechCommun.,vol.41,no.4,pp.603–623,Nov.2003.https://doi.org/10.1016/S0167-6393(03)00099-2GoogleScholarCrossref

[13] L.Chen,X.Mao,Y.Xue,andL.L.Cheng,“Speechemotionrecognition: Featuresand classification models,” Digit. Signal Process., vol. 22, no. 6, pp.1154–1160, Dec. 2012.https://doi.org/10.1016/j.dsp.2012.05.007GoogleScholarCrossref

[14] J.H.Yeh,T.-L.Pao,C.-Y.Lin,Y.-W.Tsai,andY.-T.Chen,“Segment-based emotion recognition from continuous Mandarin Chinese speech,”Comput.HumanBehav.,vol.27,no.5,pp.1545–1552,Sep.2011.https://doi.org/10.1016/j.chb.2010.10.027GoogleScholarCrossref

[15] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[16] E.M.Albornoz,D.H.Milone,andH.L.Rufiner,“Spokenemotionrecognitionusinghierarchicalclassifiers,”Comput.SpeechLang.,vol.25,no. 3, pp. 556–570, Jul. 2011.https://doi.org/10.1016/j.csl.2010.10.001GoogleScholarCrossref

[17] C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective SpeechBased on Multiple Classifiers Using Acoustic-Prosodic Information andSemanticLabels,”IEEETrans. Affect.Comput.,vol.2,no.1,pp.10–21,Jan.2011.https://doi.org/10.1109/T-AFFC.2010.16GoogleScholarCrossref

[18] S.S.Narayanan,“Towarddetectingemotionsinspokendialogs,”IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar.2005.https://doi.org/10.1109/TSA.2004.838534GoogleScholarCrossref

[19] B. Yangand M. Lugger, “Emotion recognition from speech signalsusingnewharmony features,”SignalProcessing,vol.90,no.5,pp.1415–1423,May2010.https://doi.org/10.1016/j.sigpro.2009.09.009GoogleScholarCrossref

[20] C.C.Lee,E.Mower,C.Busso,S.Lee,andS.Narayanan,“Emotionrecognitionusingahierarchicalbinarydecisiontreeapproach,”Interspeech,vol. 53, pp. 320–323, 2009.GoogleScholar

[21] S.Bjorn,S.Steidl,and A.Batliner,“TheINTERSPEECH2009EmotionChallenge,” 2009.GoogleScholar

[22] J.P. Arias,C.Busso,andN.B.Yoma,“Shape-basedmodelingofthefundamentalfrequencycontourforemotiondetectioninspeech,”Comput.Speech Lang., vol. 28, no. 1, pp. 278–294, Jan.2014.https://doi.org/10.1016/j.csl.2013.07.002GoogleScholarCrossref

[23] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0.Computers and Electrical Engineering, 104, 108481.