ISSN 2394-739X

IJRSET MARCH Volume 10 Issue 3

International Journal for Research in Science Engineering & Technology (IJRSET)

https://www.doi.org/10.5281/zenodo.8401678

Anxiety and Stress Detection through Speech Recognition using CNN

[1] C. YOSEPU
[1]Associate Professor,
[1] Department of Computer Science Engineering,
[1] St. Martins Engineering College, Secunderabad, Telangana.

ABSTRACT: Stress is a feeling of emotional tension. It canhaveaninfluenceonourmentalhealth and for the people around us. While anxiety is a natural reaction tostresswhichcanbefearfulthiscanleadtopanicattacks. These mental issues have to be addressed byeveryone.Thispaperexplainshowweareusingvocal/audiodatasettodetectstressandanxietyinaperson.Wehavedevelopedastressandanxietydetection model using deep neural network. Here audio datasets is considered from Kaggle in which the audio consists of 7 emotions i.e., joy, fear, disgust, neutral, sadness, surprised and anger. These audio datasets are used to train and test classification models like CNN. Then the audioispre-processedthroughacousticfeatureextraction, classified through CNN which provides the accuracy based on those 7emotions. By this wecanpredictif the personisstressedorhasanxiety.

Keyword: [Convolutional Neural Network, Emotion Classification, Stress Detection, MFCC (Mel frequencycepstralcoefficients), Chroma.]

1. INTRODUCTION

Thespeechrecognitionaimstodeterminetheemotionalstateofanindividualbyusinghis/ hervoice. Speechisameans of communicationtoexpress one‟s thoughts and feelings. It is one of the fast and bestways to communicate.

Speechrecognitionismostbeneficialinapplicationsthatrequire human-computer interaction such as speech syn the sis and customer service. Recognizingtheemotional state of an individual using speech signalscanbedifficultforseveralreasons.Thespeechprocessing applications has a great influence in ourlife on commercial applications like Text-to-Speechsynthesis,Speechrecognitionandverification.Speechisgivenasaninputtothemachinethataccepts this command. Theinput is translated intotextformatwhichisknownasSpeechRecognition System or Speech to Text. The speech recognitionsystemanalysesanindividualspeechinordertodetermine the emotion and produces accurate result.The speech signal which is extracted is trained byDNNmodel.Finally,theoutputobtainedwillbecomparedwithconnectedand continuousspeech.

Emotionsactasanimportantpartineachdayofhuman interactions. Emotions can be happy, fear,sad, anger and disgust. It is necessary to our rationaland intelligent decisions. It helps us to communicateanddeterminethefeelingsofotherpeoplebycommunicating our feelings and responding to others.Emotionplaysanimportantroleinshapingourbehavior.It displaystheinformationaboutmentalcondition of an individual. The speech signals can beeasily detected by using a microphone. This feature isvery useful for the users and it also helps to maintain,buildalargedatabaseforstressdetectionsystem.

Human behavior depends on the way humans act andinteract with others. Analyzing human behavior is averyimportantpracticemainlyinpsychotherapy.Behavior can be analyzed by observing the way inwhichemotionchangesduringtheconversation.Here, we implement deep learning in order to analyzetheemotionalstate.Wehavedeterminedtherelationship between emotion and behavior, furtherusedemotionstoclassifythebehaviorofanindividual. In our system, we take the input speech todetermine speech signals and then predictwhethertheindividualisunderstressor not.

2. RELATEDWORK

Recent researches for detecting Stress and Anxietyhasbeenextremelyprominent.Sometimestheresponses from Stress allows the body to overcometough situation and prepare for treats but in contrast itcan damage one‟s health too. There has been a lot ofdifferentmethodologiesfordetection of Stress and Anxietythroughphysicaltest, questionnaires that primarily rely on user input data which sometimesmay not be accurate or user may find it difficult toanswer some question if it is personal and sometimesmeasuredthrough the speech modulation and frequencythroughwhichonepersonsayshisthoughtstoothers.

In the work done by Maghilnan S, Rajesh KumarM [1] workheimplementedSentimentAnalysisbySpeechDatabyproposingfoursteps1) pre-processing which includes VAD. The input signal is givenasaninput to VAD whichidentifiesandsegregate the voice from the signal. The voices arethenstoredaschunksinthedatabase.2)SpeechRecognition System. Here the words in the languagespokenbythehumansareconvertedtomachinereadable format which is processed further. The toolsused for speech recognition are Bing speech, GoogleSpeech Recognition. 3) Speaker Recognition Systemwhere the chunks are recognized and each chunks areidentifiedandgiventheSpeakerIdithelpsinidentifying whetherthechunksarefrom the samespeakeror different. The system then matches the Speaker Id with the systemgeneratedtext. For feature extraction they haveused Mel Frequency Cepstrum Coefficient (MFCC) and for featurematchingtheyhaveused Dynamic Time Wrapping (DTW)4) Sentiment Analysis they have implemented differentalgorithmsuchasNaïveBayes, Linear SVM and VANDER and a comparisonis made to find the efficient algorithm. The accuracyfor Naïve Bayes was obtained as 72.8%, Linear SVM86.4%andVANDERas95.2%.

Kevin Tomba, Joel Dumonin , Omar Abou Khaled,Satish Hawalia [2] have discussed about multimodalstressclassificationsystemandutilizedtheaudio/video data to investigate complete number ofaudioandvideofeatureswithvariousfusiontechniquesandtemporalbackgroundsforclassificationpurposes.They showedthatTeagerenergycepstralcoefficients(TECC)surpassedstandardbaselinecharacteristicsintheaudiomodality,whilevectormodellingdependingonMFCCcharacteristicsattainedthebestprecision,while on the other hand, polynomial parameterizationoffaceimagecharacteristicsproducedthedesiredoutputacrossallsystemsandexceededthebestbaseline system. MFCCs are used as features in theextractionmodeltoextractthefeatures. Three differentdatasetswereusedtheBernilemotionaldatabaseRAVDESSdatabaseandKeioUniversityJapaneseEmotionalspeechdatabase.EmoDbandRAVDESSdatabasewereimplementedusingSVMandKeioSDdatasetusingANN.BothSVMandANNwereoptimizedwiththehelpofScikit-learnlibrarymethod.Thismethodwasusedinfindingthebestcombinationofvaluestogivethebestresultforasetoffeatures.SVMandneuralnetworkswereusedintheclassification.Bothalgorithmsshowedbestresults,withANNshavingslightlybetterscoresthanSVMs.Theobtainedresultsperformedgoodclassification and determined if there is stress or not.AndersonR.Avila,ShrutiR.Kshirsagar,AbhishekTiwari,DanielLafond,DouglasO‟ShaughnessyandTiagoH.Falk[3]heusedaCNN,SVMandDNNlearningtechniquesandevaluatedwhichmodelyieldsthehighestaccuracy.ThisexperimentwasperformedusingSpeechUnderStimulatedandActualStress(SUSAS)dataset.They proposedtheuse ofmodulationspectralfeatures (MSF) as aninputtoCNNandadoptedOpenSMILEfeaturesandevaluateditwithSVMandDNN.Inordertoextractmodulationspectralfeaturesthespeechsignalisfirstnormalizedto-26dBovandeliminatingunwantedspeechsignals.Thentheyhavefilteredthesignalsusingkmodulationfilterslaterthefrequenciesfromthefiltercenterareequallyspacedfrom4to128Hz.Finallyfivefeaturesetareextracted.TheresultsshowedthattheproposedMSFcombinedwithCNNoutperformedtheothertwolearningmethodsSVMandDNNandgaveanoverallaccuracy of72%whileDNNmethodachieved62%accuracyandSVMproduced61%accuracy.

Dr.S.Vaikole, S. Mulajkar, A. More, P. Jayaswal, S.Dhas[4]proposedanalgorithmthatfirstextractsMel-filterbankcoefficientsusingapre-processedspeech data and then predicts the stress output usingCNN.Theaudiosignalispassedtospeechpreprocessingandthenforwardedtofeatureextraction module. All the necessary speech featuresare extracted and are passed to a deep-learning basedstress detection model. The CNN model determinestheuser‟sstressstatebyadecisionprocess.TheproposedsystemusesRavdessdatabase.Totalof1440 Speech utterances of twelve male and femalespeakerswere taken. Labelswere usedfor trainingthemodelusingone-hot-encodingapproach.TheaccuracywasclassifiedintopitchrateandMFCC. Theproposedmodelconsistsof eightCNNlayersand fully connected layers. These layers capture thenecessary information of extracted features and thencalculatetheframe-leveloutputeachtime.Theoutput of frame-level is converted into a sentence-levelfeature. The features extracted from layers areof two types that is average value of output sequenceand last frame-level output. The accuracy of stressdetection system using pitch rate was 52% and usingMFCCwas 94.33%. He furtherconcludedthat byusing signal raw energy operator stressed emotionsaredetectedwithimprovedaccuracy.

Arushi, Roberto Dillon, Ai Ni Teoh [5] proposed aVR-based stress detection model where the speaker‟svoice is analyzed on real-time basis where virtuallythe speaker‟s speaking skills to the audience will beimprovedbyreal-insightsfromthegamewhichprovides the support/feedback. They have taken thedataset Ryerson Audio-Visual Database of EmotionalSpeechandSong(RAVDESS).Theyhaveconstructed 3 classifiers models to extract the voicefeaturesAmplitudeEnvelope(AE),Root-Mean-Square(RMS)andMel-FrequencyCepstralCoefficients (MFCCs). Using Random Forest, KNN& SVM training and testing of data is done. MachinelearningAlgorithmslikeGaussianMixtureModel(GMM), Hidden Markov Model (HMM), ArtificialNeural Networks (ANN) and Deep Neural Network(DNN).VRbasedstressdetectionmodelincludesvirtualenvironment,behaviorofvirtualaudience,machinelearningmodeldevelopment,featureselection, training and testing of model development.In this model they have kept 70% of data for trainingand 30% for testing the actor‟s voice dataset.

Thefinalresultsshowsthatrandomforestaccuracyis82%, KNN accuracy is 72% and SVM accuracy 57%,5% and24% accuracy has increasedtodetectthestresswithfeaturesthatincludesRMS,AEandMFCC.

3. METHODOLOGY

Dataset Collection

Theactorbasedspeechdatabaseiscomprisedof2768files.Onem

ISSN 2394-739X
IJRSET MARCH Volume 10 Issue 3

otionalvalidity,strength,andgenuineness, each filewas scored 10 times. Therewere 24 individuals that were characterized by an un-trainedadultstudycandidatesbelongingtoNorthAmerica were given scores. High emotional validitylevels,reliabilityofinterrater,andreliabilityoftest-retest interrater were recorded. In the database, thereare 24 trained actors (12 male, 12 female), in a NorthAmericanneutralvoice,clearlyexpressingtwolinguisticallyrelatedphrases.Speechincludesexpressionsofneutral,happy,sad,angry,fear,disgust surprise and calm. At two emotional intensityratios,(strongandnormal),eachexpressionisgeneratedwithanadditionalneutralexpression.Therearethree modeformatsavailableforallconditions:audio-only(16bit,48kHz.wav).

Speech Recognition

Speech recognition is the way of converting acoustics(speech of a person) into textual form. This is widelyused in virtual assistants like Rebecca, Siri, Alexa,etc.ThegoogleAPIcalledSpeechRecognitionwhich allows us to convert speech into textual forfurtherprocessingbutwhileusingtheSpeechRecognition API, translating big or long audio filesinto text, it may give error messages because it is notthatstrongforlargefilesofaudio.Firstly,weinternally see the input physical audio which will getconverted into electric signals. The electric signals ofour speech signal then gets converted into digitizedform with an analog-to-digital converter. Then, thedigitized model can be used to transcribe the speechinto textualform.

Feature Extraction

Acoustic Features: In general, the more precise andvery basic features of audio to recognize affect areconsidered to be duration, MFCC, energy and pitch.This has been supportedby a many researchworkand found it to be themost correct acousticfeaturesto emotions are duration and energy, while all theother featuresareofmediumrelevance.

Mel-FrequencyCepstralCoefficients(MFCC)depending on a linear cosine transformation (CT) of alog power spectrum performed on a non-linear Melfrequency scale, it is known as the spectrum of short-term control of an audio or sound. Any type of soundcreatedbyhumansisdefinedbytheirvocaltractshape, including tongue, teeth, lips, etc. The envelopeof the time power spectrum of the audio signal isrepresentative of the vocal tract and MFCC, definedas the coefficients that make up the Mel-frequencycepstrumandcorrectlyrepresentthisenvelope. Options are considered for the lower dimensions ofthe1stthirteenMFCCcoefficientsastheyrepresent thespectraenvelope.Anditsspectraldataisindicatedbythe higherdimensionswhicharediscarded.Envelopesarenecessaryfordifferentphonemes to display the difference, so we can findphonemesthroughMFCC.Chroma:Itisalso called as „Chromagram‟,„Pitchclassprofiles‟,„Chromafeatures‟, that relates to the twelve different kinds ofpitch classes and tuning approximated to the equaltemperedscale. It basically computesmelodic andharmonic characteristics of speech or an audio signal.Itisconsistingof2features:

Chroma Vector: It has twelve element expression of spectral energy.

Chromadeviation: It isthetwelve Chromaparametersstandarddeviation.

Convolutional Neural Network

Thedeeplearningmodeldependingupontheconvolutionary neural network (CNN) is used and itsdenselayershavebeenused.

Astheonlyaudiofeaturetotrainour CNN model, the MFCC and Chroma features are considered the basic approach. The MFC Ccoefficientswereonlyusedfortheirability to reproduce the amplitude spectrum of the audio wave in a compact vector form. As mentioned in, the speech file is split into frames, using a fixedwindowsize.

The discrete Fourier transform is implemented, then the logarithm of the amplitude spectrum is taken into account. After a certain amount of frequency 'Mel'reduction, the spectrumofamplitudeisthennormalized. For a significant re-construction of the sound wave that can be distinguished by the humanauditoryprocess, this techniqueisperformed to empathize the frequency to a more realistic type. For eachspeechfile, somefeatureswereextracted. Features were produced and along with it converting each speech file to a time series of floating points Then MFCC sequencewascreatedfromthetimeseries.

If the input given is a size < set of training samples >x n x 1 onwhich we executed a one-dimensional CNN round as the activation function Re Lu and 2 x2isthemax-poolingfunction. ReLuasg (z)=max {0,z},anditgetsalargevalueinthecaseofactivationby addingthisfunctiontorepresentthehidden units. The last activation layer is used as theSoftmax layer which calculates relative probabilities.Thenattheendthefullyconnectedlayerisused where the classification happens. Pooling allows theCNN model to focus only on the main characteristicsof each of the data components, not segregating themby their position. The output of the pooling layer isflattened and this flattened matrix is fed into the fullyconnected layer.

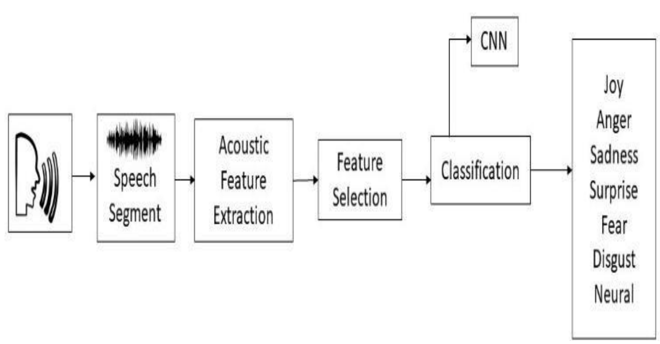

Figure 1: System Overview

4. RESULT

Thefindingsattainedfromtheevaluationprocessindicatetheeffi cacyofthemodelonthedatasetrelative to the baselines and the state of the art. Itshows the precision, recall and F1 score values thatwere attained for each of the emotional groups. Thesefindings suggest that recall and accuracy are kind ofbalanced, enabling us to achieve a 0.76 F1 score fortheclass.TheslightshiftinF1highlightstherobustness of the CNN model, which manages 76.08percentaccuracyeffectively.

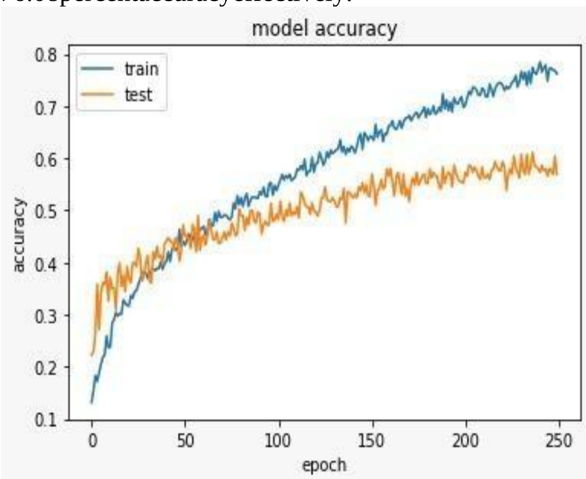

Figure 2: CNN Model Epoch

CONCLUSION

This work presents a deduced model that takes audio as an input and identifies whether the user is under Stress and Anxiety. In this paper we have proposed asimplesystemtocarryouttheabovementionedfunctions. Wehaveextracted the MFCC, MEL and Chromogramfeaturesfromtheaudiofilesusedthroughouttrainingtoacquiresuchresults.Wetrainedourneuralnetworkontheaboverepresentations of inputdatatocorrectly figureoutthe probability of distribution of annotation sectionsemploying1-DimensionalCNN,max-poolingandDense Layers. The result gained can only be worth itas astarting pointforfurtherexpansions,updates,and enhancements.

REFERENCES

[1] Maghilnan S,Rajesh KumarM-"Sentiment Analysison Speaker Specific Speech Data",2017 International Conferenceon Intelligent Computing and Control (I2C2).

[2]. Kevin Tomba, Joen Dumoulin, Elena Mugellini,Omar Abou Khaled and Salah Hawila-"Stressdetectionthroughspeechanalysis",2018.

[3]. B.Padmaja,V.V.Rama Prasad and K.V.N.Sunitha" A Machine Learning Approach for Stress Detectionusinga Wireless Physical Activity Tracker",2018.

[4]. Anderson R.Avila, Shruti R.Kshirsagar, Abhishek Tiwari, Daniel Lafond, DouglasO"Shaughnessy and TiagoH.Falk-"Speech-BasedStress Classification Based On Modulation Spectral Features And Convolutional Neural Networks", 2019 27th European SignalProcessing Conference (EUSIPCO).

[5]. Ravinder Ahuja, Alisha Banga – "Mental StressDetection in University Students using Machinelearning", 2019.

[6]. Zhao Cheng Huang, Julien Epps, Dale Joachim,Vidhya Saharan Sethu, IEEE Journal of SelectedTopics in – "SignalProcessing, Natural LanguageProcessingMethodsinSpeech-basedStressDetectionforAcousticandLandmarkEvent-based Features", 2019.

[7]. RussellLi, Zhidong Liu-"Stress Detection Using DeepNeural Networks",2020.

[8]. Dr.S.Vaikole,S.Mulajkar,A.More,P.Jayaswal,S.Dhas-"StressDetectionthroughSpeechAnalysisusingMachineLearning",Volume8, Issue 5May2020.

[9]. Zarinatj Hosseinzadeh-Shanjanji, KhadijehHajiMiri, BahramRostami, ShokoufehRamezani, Mohsen Dadshi,-"StressAnxietyLevelsAmong Healthcare Staff During COVID-19 Epidemic",2020.

[10]. Arushi, Roberto Dilloa and Ai Ni Teoh,"Real-time stress detection model and voice analysis:An integrated VR-based game for training publicspeaking skills", 2021.

[11]. G. Jawaherlalnehru S. Jothi Shri, S. Jothilakshmi, "Real Time Face Recognition in Group Images using LBPH"International Journal of Recent Technology and Engineering (IJRTE), 2019.