WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Emotion Recognition System Using Machine Learning

Husbaan I. Attar, Nilesh K. Kadole, Omkar G. Karanjekar, Devang R. Nagarkar, Prof. Sujeet More

Bachelor of Engineering,Department of Information Technology, Trinity College of Engineering& Research (TCOER), Pune

ABSTRACT: -

Affective Computing Aims at Providing Effective and Natural Interaction Between Human and Computers. One Important Goal Is to Enable Computers to Understand the Emotional States Expressed by The Human Subjects, So That Personalized Responses Can Be Delivered Accordingly. Most Of Studies in The Literature Are Focused on Emotion Recognition from Isolated Short Sentences, Which Hinders It from Practical Applications. In This Chapter, We Explore Emotion Recognition from Continuous Speech and Propose a Real-Time Speech Emotion Recognition System. The System Consists of Voice Activity Detection, Speech Segmentation, Signal Pre-Processing, Feature Extraction, Emotion Classification, And Statistics Analysis of Emotion Frequency. The Experiments with Both Pre-Recorded Datasets and Real-Time Recording Expressed in Four Different Emotion Categories Have Been Carried Out. The Average Accuracies Of 90% And 78.78% Are Achieved in The Two Experiments, respectively. We Also Investigate the Application of The Developed Real-Time Speech Emotion System In Online Learning. The Results from The Experiment In A Simulated Online Learning Environment Show That Our Emotion Recognition System Can Efficiently Recognize The Student's Response To The Course. This Enables Online Courses To Be Customized To Fit Students With Different Learning Abilities And Helps Students To Achieve Optimal Learning Performance.

1.Introduction

Speech Is An Important Carrier Of Emotions In Human Communication. Speech Emotion Recognition (SER) Has Wide Application Perspectives On Psychological Assessment, Robots, Mobile Services, Etc. The Emotions Of A Person Influence Various Physical Aspects Like Muscular Tension, Skin Elasticity, Blood Pressure, Heart Rate, Breath, Tone Of Voice Etc. Some Of These Physical Reflections Of Emotions Are Much More Obvious And Externally Accessible Than Others, Like The Expression And Mimic Of The Face, The Tone And Pitch Of The Voice.In Order To Communicate Effectively With People, The Systems Need To Understand TheEmotions In Speech. Therefore, There Is A Need To Develop Machines That Can Recognize The Paralinguistic Information Like Emotion To Have Effective Clear Communication Like Humans.

A Lot Of Machine Learning Algorithms Have Been Developed And Tested In Order To Classify These Emotions Carried By Speech. The Aim To Develop Machines To Interpret Paralinguistic Data, Like Emotion, Helps In Human-Machine Interaction And It Helps To Make The Interaction Clearer And Natural. In This Study Convolution Neural Networks Are Used To Predict The Emotions In Speech Sample.

1. Software Requirements

- Programming Language - Python
- ANACONDA 3-64bit
- Jupyter Notebook
- Operating System - Any OS Like A Window, Ubuntu.

Kit Required To Develop Speech Emotion Recognition Using Python:

- No Kit Required

Technologies You Will Learn By Working On Speech Emotion Recognition Using Python:

- Python

2. Problem Statement

- Often in the interest to increase the acceptability of speech technology for human users. The speech signal communicates linguistic information between speakers as well as paralinguistic information about speaker’s emotions, personalities, attitudes, feelings, levels of stress and current mental states.
- Words are not enough to correctly understand the mood and intention of a speaker and thus the introduction of human social skills to human-

machine communication is of paramount importance. This can be achieved by the researching and creating methods of speech modelling and analysis that embrace the signal, linguistic and emotional aspects of communication.

- It can be used to improve the robustness of speech and speaker recognition systems. Moreover, by assessing a speaker's speech, emotion classification can support automatic assessment of mental states of people working in dangerous environments (e.g. chemicals, explosives) and people undertaking high levels of responsibility (e.g. pilots, surgeons).
- all systems can use emotion recognition to sort emergency telephone messages, or cope with disputes through monitoring the mental states (levels of satisfaction) of customers. Another commercial application of emotion detection system is the interactive game industry offering the sensation of naturalistic human-like interaction to player's mood as well as the ability to respond accordingly through affective voice or face expression.

3. Existing System

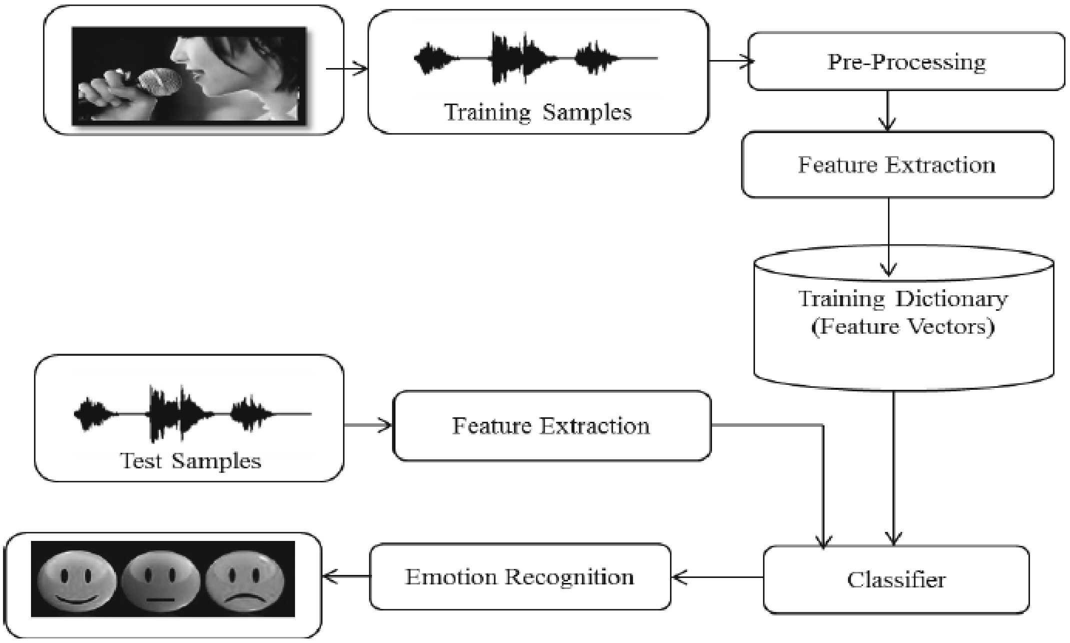


Fig 6.1 Architecture

Modules:

In Our CNN Model We Have Four Important Layers:

1. Convolutional Layer: Identifies Salient Regions At Intervals, Length Utterances That Are Variable And Depicts The Feature Map Sequence.
2. Activation Layer: A Non-Linear Activation Layer Function Is Used As Customary To The Convolutional Layer Outputs. In This We Have Used Corrected Linear Unit (Relu) During Our Work.
3. Max Pooling Layer: This Layer Enables Options With The Maximum Value To The Dense Layers. It Helps To Keep The Variable Length Inputs To A Fixed Sized Feature Array.
4. Dense Layer

4. Existing System Algorithm

CNN Algorithm:

//Anaconda With Jupyter Notebook Tool In Python Language.

Step 1: The Sample Audio Is Provided As Input.

Step 2: The Spectrogram And Waveform Is Plotted From The Audio File.

Step 3: Using The LIBROSA, A Python Library We Extract The MFCC (Mel Frequency Cepstral Coefficient) Usually About 10–20.

//Processing Software

Step 4: Remixing The Data, Dividing It In Train And Test And There After Constructing A CNN Model And Its Following Layers To Train The Dataset.

Step 5: Predicting The Human Voice Emotion From That Trained Data (Sample No. - Predicted Value - Actual Value)

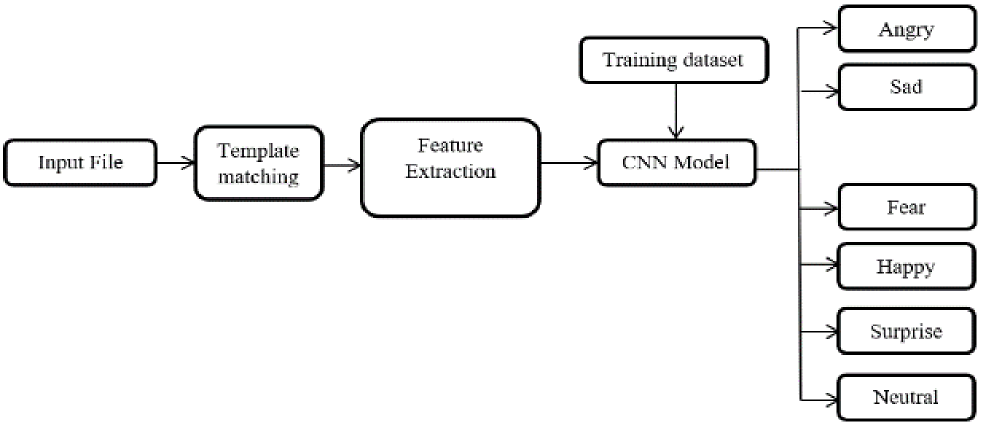


Fig 6.2 CNN Algorithm

5. System Implementation

Methodology Used

In this work, we conduct an extensive comparison of various approaches to speech based emotion recognition systems. The analyses were carried out on audio recordings from Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) . After pre-processing the raw audio files, features such as Log-Mel Spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs), pitch and energy were considered. The significance of these features for emotion classification was compared by applying methods such as Long Short Term Memory (LSTM) , Convolutional Neural Networks (CNNs), Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs).

Methodology:

The speech emotion recognition application is executed using convolutional neural network. Following is the architecture of the system:

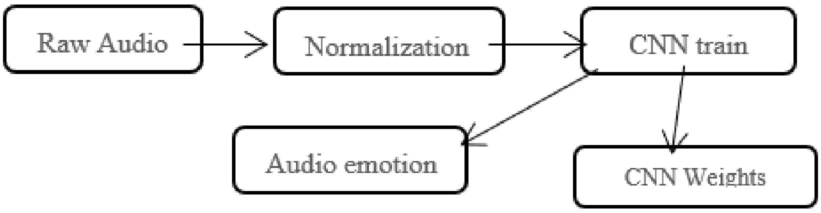


Fig 6.1 Methodology

Training Model and Testing Model:

A training data is fetched to the system which consists the expression label and Weight training is also provided for that network. An audio is taken as an input. Thereafter, intensity normalisation is applied over the audio. A normalised audio is used to train the Convolutional Network, this is done to ensure that the impact of presentation sequence of the examples doesn't affect the training performance. The collections of weights come out as an outcome to this training process and it acquires the best results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final network weights trained it gives the determined emotion. The output is represented in a numerical value each corresponds to either of five expressions.

There are 3 emotions that are being detected based on the person's bpm value, those are Relaxed/Calm, Joy/Amusement, Fear/Anger. The produced art's colors and shapes are parallel to the detected emotion based on the principles of "color psychology" and "shape psychology".

Speech Database:

In this survey different speech database are utilized to validate the proposed methods in speech emotion recognition. Among all dataset Berlin and AIBO are most common used. Burkhardt et al. was recorded by actors in German language. The place of record was Department of Technical Acoustics of Technical University Berlin. 5 male and 5 female German actor have participated in providing the dataset by reading one of the chosen sentences. Different recorded emotion are anger, fear, neutral, disgust, happiness and sadness. Another emotional database names Aibo was collected in the real conditions by interacting and playing of fifty-one children with the Sony's robot Aibo that govern by human operator to extract the children's spoken speech. In AIBO five collected emotions are positive, neutral, angry, rest and emphatic.

Feature Extraction:

The next step involves extracting the features from the audio files which will help our modellearn between these audio files. For feature extraction we make use of the LibROSA library inpython which is one of the libraries used for audio analysis.When we do Speech Recognition tasks, MFCCs are the state-of-the-art feature since it was invented in the 1980s.

This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

Classification Approaches:

For modeling the emotional states, there are different classification methods utilized to create proper classifier such as support vector machine (SVM), hidden Markov models (HMM), neural network, Knearest neighbor and Gaussian mixture model (GMM). Conversely, a standard level of classifier may not achieve on very emotional statuses. For example ranking SVM approach cannot leads to considerable improvements in recognition of emotion compare to combination of SVM with radial basis function (RBF). Some hybrid/fusion-based methods achieve high recognition rate compare to individual approaches.

Waveform and Spectrogram:

We tested out one of the audio files to know its features by plotting its waveform andspectrogram.

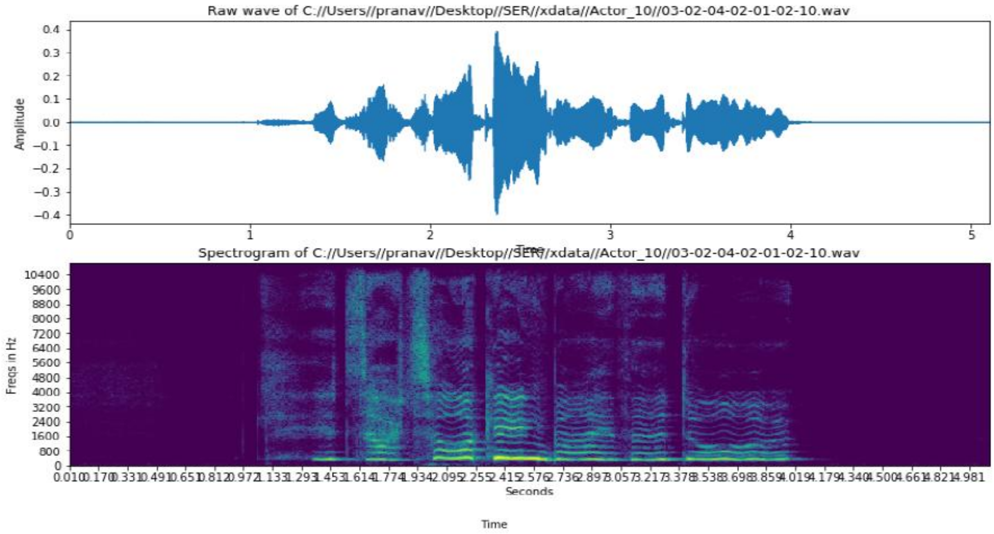


Fig 6.2 Waveform and Spectrogram

We would use MFCCs to be our input feature. If you want a thorough understanding of MFCCs. Loading audio data and converting it to MFCCs format can be easily done by the Python package librosa.

Required Packages:

Table 6.1 Required Packages

No.	Package Name	Description	Version
1.	librosa	Python package for music and audio analysis.	0.8.1
2.	tensorflow	Tensorflow is a machine learning library.	2.3.0
3.	keras	Deep learning library for theano and thesorflow.	2.4.3
4.	pandas	High-performance, easy to use data structures and analysis tools.	1.3.4
5.	wave	Python Wand ImageMagick library which is used to alter an image along with a sine wave. It creates a ripple effect.	0.0.2
6.	pyaudio	Bindings for portaudio v19, the cross-platform audio stream library.	0.2.11
7.	pylint	Python code static checker.	2.12.2
8.	scikit-learn	A set of python modules for machine learning and data mining.	1.0.1
9.	glob2	Version of the glob module that supports recursion via**, and can capture pattern.	0.7
10.	ipython	Productive interactive computing.	7.29.0
11.	matplotlib	Publication quality figures in python.	3.5.0
12.	numpy	Arry processing for numbers, strings, records and objects.	1.21.2
13.	scipy	Scientific library for python.	1.7.1
14.	seaboarn	Statistical data visualization.	0.11.2
15.	tqdm	A fast, extensible progress meter.	4.62.3
16.	plotly	An interactive javascript-based visualization library for python.	5.1.0

Data Set

Table 6.2 RAVDESS Dataset Actors Voice

Actors	Actor Recorded Voise Count
Actor 1	60
Actor 2	60
Actor 3	60
Actor 4	60
Actor 5	60
Actor 6	60
Actor 7	60
Actor 8	60
Actor 9	60
Actor 10	60
Actor 11	60
Actor 12	60
Actor 13	60
Actor 14	60
Actor 15	60
Actor 16	60
Actor 17	60
Actor 18	60
Actor 19	60
Actor 20	60
Actor 21	60
Actor 22	60
Actor 23	60
Actor 24	60

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

- 12 Actors & 12 Actresses recorded speech and song versions respectively.
- Actor no.18 does not have song version data.
- Emotion Disgust, Neutral and Surprised are not included in the song version data.

We went with the Audio only zip file because we are dealing with finding emotions from speech.The zip file consisted of around 1500 audio files which were in wav format.

The second website contains around 500 audio speeches from four different actors with different emotions.The next step involves organizing the audio files. Each audio file has a unique identifier at the 6th position of the file name which can be used to determine the emotion the audio file consists of.We have 5 different emotions in our dataset.

1. Calm
2. Happy
3. Sad
4. Angry
5. Fearful

- We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Using the librosa library we were able to extract features i.eMFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated out the females and males voices by using the identifiers provided in the website. This was because as an experiment we found out that separating male and female voices increased by 15%. It could be because the pitch of the voice was affecting the results.
- Each audio file gave us many features which were basically an array of many values. These features were then appended by the labels which we created in the previous step.
- The next step involved dealing with the missing features for some audio files which were shorter in length. We increased the sampling rate by twice to get the unique features of each emotional speech. We didn’t increase the sampling frequency even more since it might collect noise thus affecting the results.

Experimental Setup

This section contains explanations about experimental setup, libraries used for Speech Emotion Recognition which helps in emotion recognition.

A. SYSTEM SETUP

For performing the experiment I’ve used a system setup consisting of Core i7 7th Generation, 8GB RAM, System type: Windows 10 64-bit operating system. For deep learning I’ve used TensorFlow for implementing the Inception net model and Tensor Board for visualizing the learning, graphs, histograms and so on.

B. TRAINING METHOD

All images labelled with respective emotions are prepared for training the model. The proposed CNN model was implemented using TensorFlow. The spectrogram images generated from the RAVDESS are resized. Spectrograms were generated from all the audio files in the dataset.

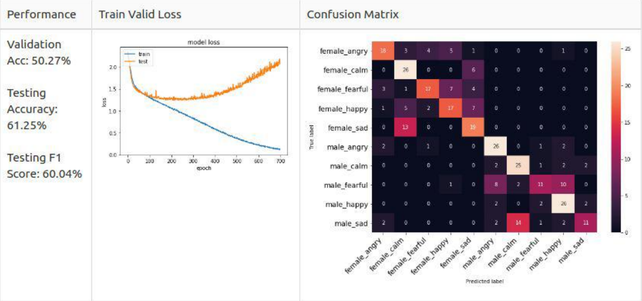
Exploratory Data Analysis:

In the RADVESS dataset, each actor has to perform 4 emotions by saying and singing twosentences and two times for each. As a result, each actor would induce 4 samples for each emotion except neutral, disgust and surprised since there is no singing data for these emotions. Each audio wave is around 4 second, the first and last second are most likely silenced.

Replicating Result:

As they excluded the class neutral, disgusted and surprised to do a 10 class recognition for theRAVDESS dataset.

I tried to replicate his result with the model provided, I can achieve a result of



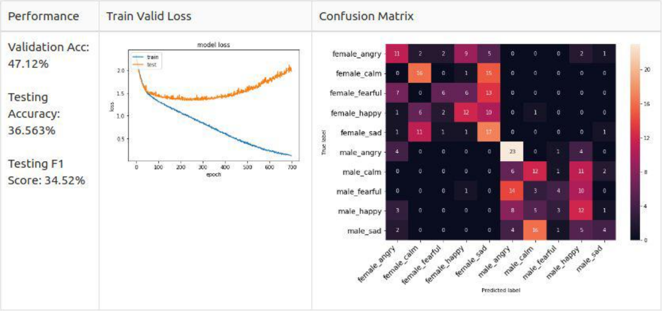
Performance	Train Valid Loss	Confusion Matrix
Validation Acc: 50.27% Testing Accuracy: 61.25% Testing F1 Score: 60.04%		

Fig 6.3 Replicating Result (1)

However, I found out there is a data leakage problem where the validation set used in the trainingphase is identical to the test set. So, I re-do the data splitting part by isolating two actors and two actresses data into the test set which makes sure it is unseen in the training phase.

- Actor no. 1–20 are used for Train / Valid sets with 8:2 splitting ratio.
- Actor no. 21–24 are isolated for testing usage.
- Train Set Shape: (1248, 216, 1)
- Valid Set Shape: (312, 216, 1)
- Test Set Shape: (320, 216, 1) — (Isolated)

I re-trained the model with the new data-splitting setting and here is the result:



Performance	Train Valid Loss	Confusion Matrix
Validation Acc: 47.12% Testing Accuracy: 36.563% Testing F1 Score: 34.52%		

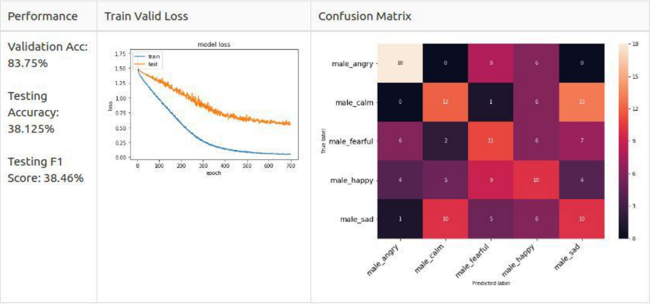
Fig 6.4 Replicating Result (2)

Benchmark:

From the train valid loss graph, we can see the model cannot even converge well with 10 target classes. Thus, I decided to reduce the complexity of my model by recognizing male emotions only. I isolated the two actors to be the test set, and the rest would be the train/valid set with 8:2 Stratified Shuffle Split which ensures there is no class imbalance in the dataset. Afterward, I trained both male and female data separately to explore the benchmark.

Augmentation

After I tuned the model architecture, optimizer and learning rate schedule, I found out the model still cannot converge in the training period. I assumed it is the data size problem since we have 800 samples for train valid set only. Thus, I decided to explore the audio augmentation methods. Let’s take a look at some augmentation method with code. I simply augmented all of the datasets once to double the train / valid set size.



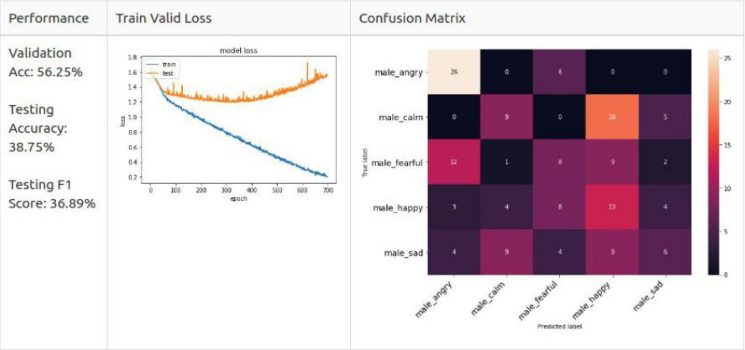
Performance	Train Valid Loss	Confusion Matrix
Validation Acc: 83.75% Testing Accuracy: 38.125% Testing F1 Score: 38.46%		

Fig 6.5 Augmentation

Male Dataset

Train Set = 640 samples from actor 1- 10.
Valid Set = 160 samples from actor 1- 10.
Test Set = 160 samples from actor 11- 12.

Male Baseline



Performance	Train Valid Loss	Confusion Matrix
Validation Acc: 56.25% Testing Accuracy: 38.75% Testing F1 Score: 36.89%		

	male_angry	male_calm	male_fearful	male_happy	male_sad
male_angry	26	0	6	0	0
male_calm	0	9	0	18	5
male_fearful	12	1	8	9	2
male_happy	3	4	8	13	4
male_sad	4	9	4	9	6

Fig 6.6 Male Baseline

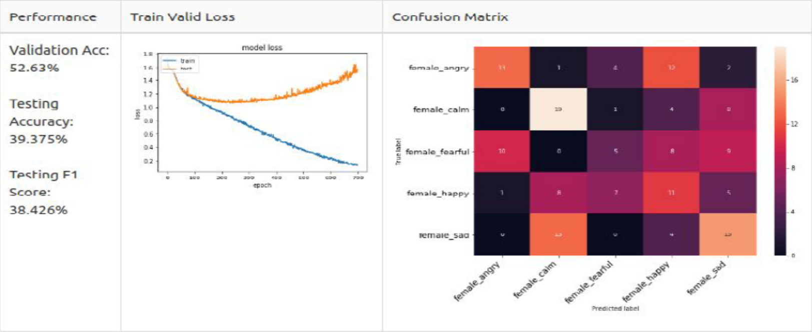
Female Dataset

Train Set = 608 samples from actress 1- 10.

Valid Set = 152 samples from actress 1- 10.

Test Set = 160 samples from actress 11- 12.

Female Baseline



Performance	Train Valid Loss	Confusion Matrix
Validation Acc: 52.63% Testing Accuracy: 39.375% Testing F1 Score: 38.426%		

	female_angry	female_calm	female_fearful	female_happy	female_sad
female_angry	13	1	4	12	2
female_calm	0	19	1	4	8
female_fearful	10	0	5	8	9
female_happy	1	8	7	11	5
female_sad	0	13	0	4	15

Fig 6.7 Female Baseline

Results Accuracy

In this Python project, we learned to recognize emotions from speech. We used an MLPClassifier for this and made use of the sound file library to read the sound file, and the librosa library to extract features from it.As you'll see, the model delivered an accuracy of 80.00%. That's good enough for us yet.

Accuracy Score: 80.00%

6. Implementation Diagram

Use Case Diagram

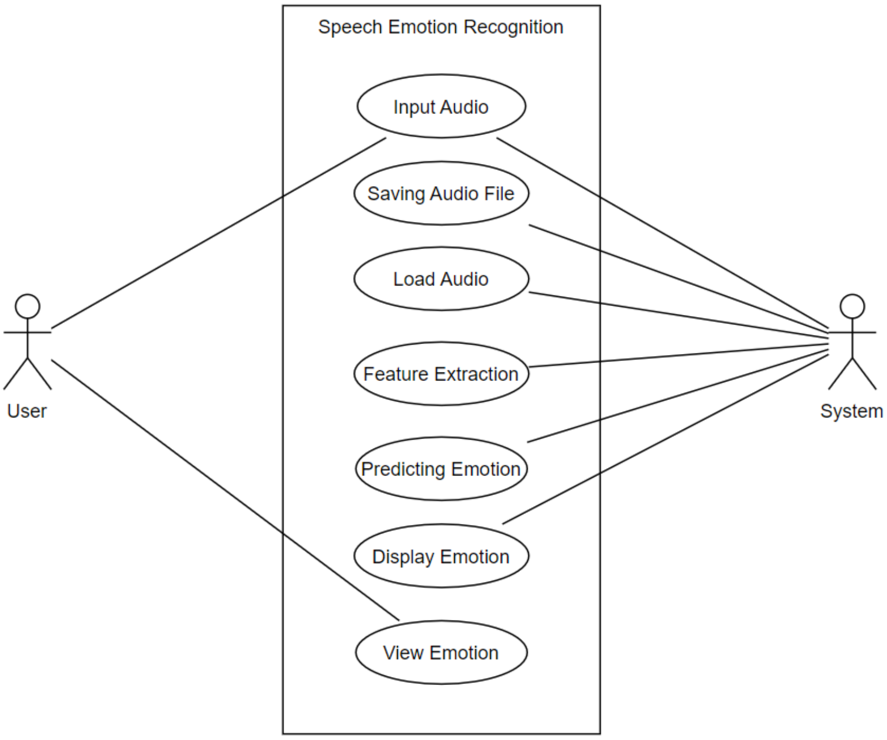


Fig 7.1 Use Case Diagram

Activity Diagram

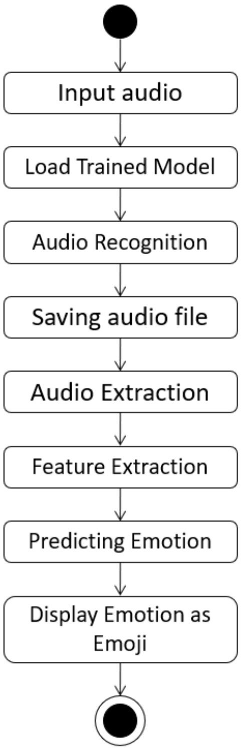


Fig 7.2 Activity Diagram

Flow Chart

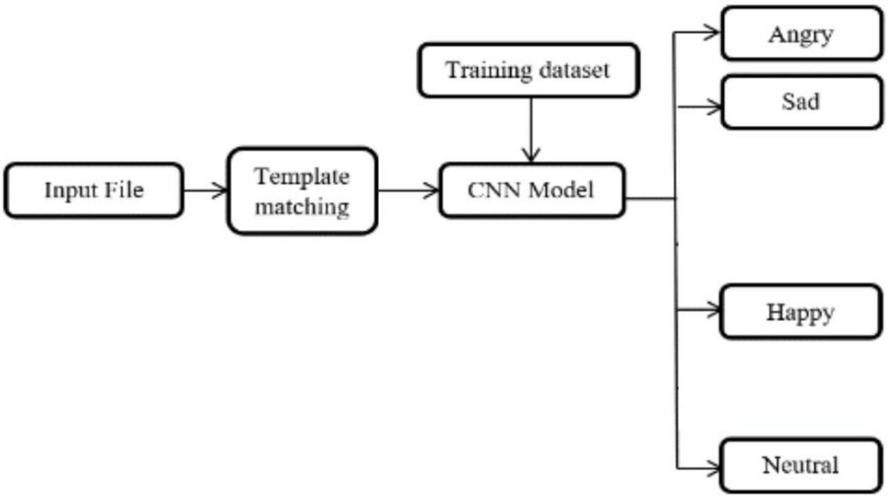
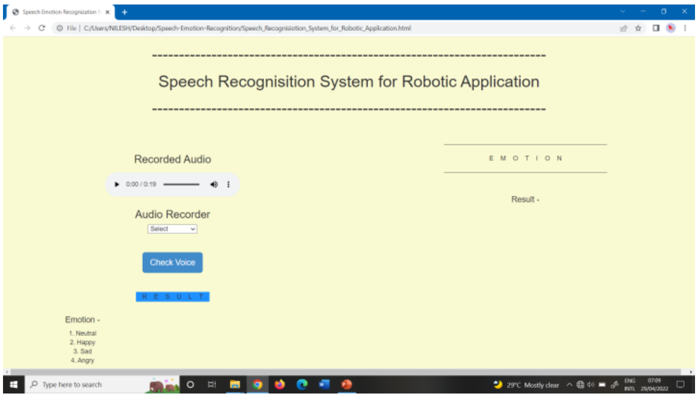


Fig 7.3 Flow Chart

Output


Speech Emotion Recognization f
File | C:/Users/NILESH/Desktop/Speech-Emotion-Recognition/Speech_Recognisiotion_System_for_Robotic_Application.html
0:00 / 0:19
Select
Check Voice
R E S U L T
Type here to search
29°C Mostly clear
ENG INTL 07:09 29/04/2022

Speech Recognisition System for Robotic Application

Recorded Audio

Audio Recorder

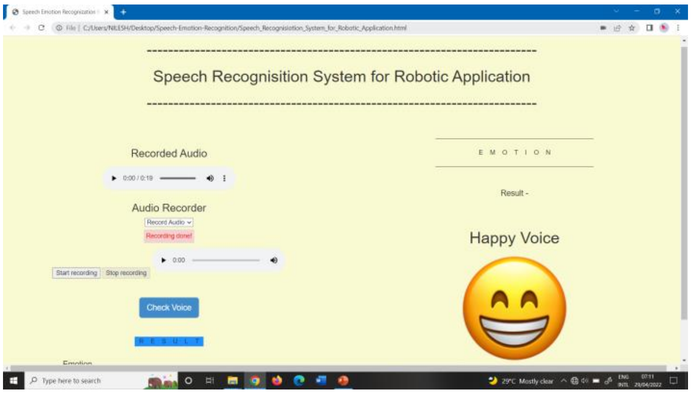
E M O T I O N

Result -

Emotion -

1. Neutral
2. Happy
3. Sad
4. Angry

Fig 8.1 Sample GUI


Speech Emotion Recognization f
File | C:/Users/NILESH/Desktop/Speech-Emotion-Recognition/Speech_Recognisiotion_System_for_Robotic_Application.html
0:00 / 0:19
Record Audio
0:00
Start recording
Stop recording
Check Voice
R E S U L T
Type here to search
29°C Mostly clear
ENG INTL 07:11 29/04/2022

Speech Recognisition System for Robotic Application

Recorded Audio

Audio Recorder

Recording done!

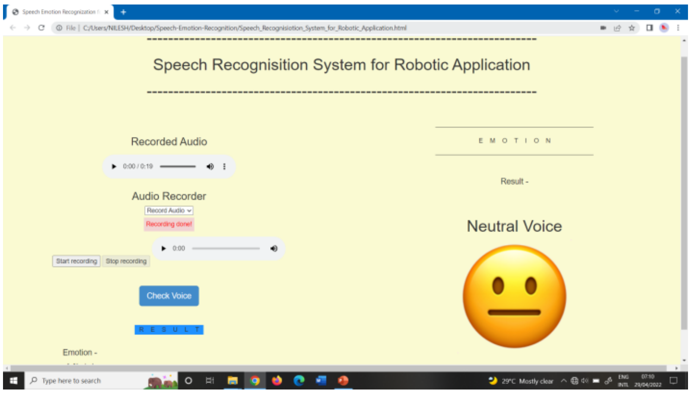
E M O T I O N

Result -

Happy Voice

Emotion

Fig 8.2 Happy Emotion



Speech Recognisition System for Robotic Application

Recorded Audio

Audio Recorder

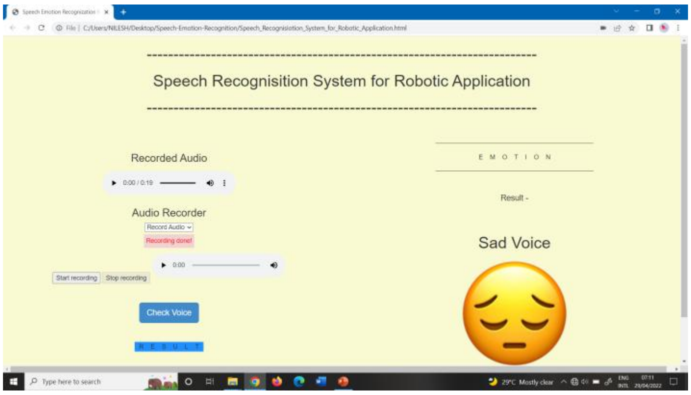
E M O T I O N

Result -

Neutral Voice

Emotion -

Fig 8.3 Neutral Emotion



Speech Recognisition System for Robotic Application

Recorded Audio

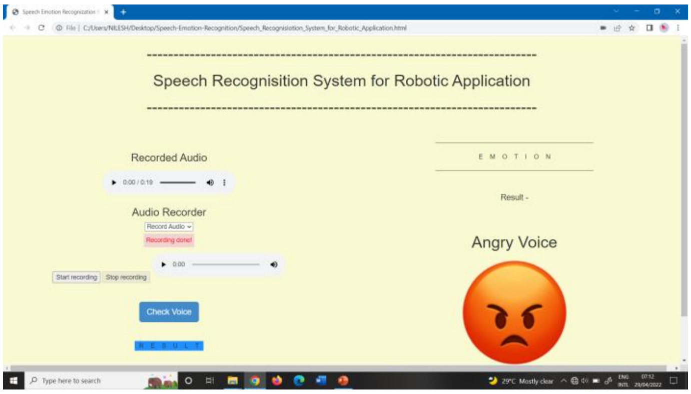
Audio Recorder

E M O T I O N

Result -

Sad Voice

Fig 8.4 Sad Emotion



Speech Recognisition System for Robotic Application

Recorded Audio

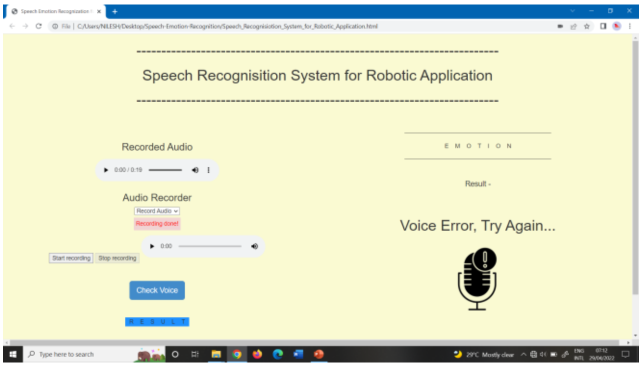
Audio Recorder

E M O T I O N

Result -

Angry Voice

Fig 8.5 Angry Emotion



Speech Recognisition System for Robotic Application

Recorded Audio

Audio Recorder

E M O T I O N

Result -

Voice Error, Try Again...

Fig 8.6 Voice Error

Conclusion

Our Project Can Be Extended To Integrate With The Robot To Help It To Have A Better Understanding Of The Mood The Corresponding Human Is In, Which Will Help It To Have A Better Conversation As Well As It Can Be Integrated With Various Music Applications To Recommend Songs To Its Users According To His/Her Emotions, It Can Also Be Used In Various Online Shopping Applications Such As Amazon To Improve The Product Recommendation For Its Users.

References

[1] Mohammed AbdelwahabAnd Carlos Busso, Multimodal Signal Processing (MSP) Laboratory, Erik Jonsson School OfEngineering &Computer Science, University Of Texas At Dallas, Richardson, Texas 75083, U.S.A.

[2] Voice Based Emotion Recognition WithConvolutional Neural Networks For Companion Robots Eduard FRANT¸ I1, 2, Ioan ISPAS1, Voichita DRAGOMIR3, Monica DASCAˇ LU1, 3, Elteto ZOLTAN1, And Ioan Cristian STOICA4,

[3] MULTIMODAL SPEECH EMOTION RECOGNITION USING AUDIO AND TEXT Seunghyun Yoon, Seokhyun Byun, And Kyomin Jung Dept. Of Electrical AndComputer Engineering, Seoul National University, Seoul, Korea Fmysmilesh, Byuns9334, Kjungg@Snu.Ac.Kr

[4] Speech Emotion Recognition Using CNN, Article InInternational Journal Of Psychosocial Rehabilitation · June 2020, Harini Murugan

[5] Speech Emotion Recognition Methods: A Literature Review Babak Basharirad, And Mohammadreza Moradhaseli

[6] SPEECH EMOTION RECOGNITION Darshan K.A1, Dr. B.N. Veerappa2 1U.B.D.T. College OfEngineering, Davanagere, Karnataka, India 2Dr. B.N. Veerappa, Department Of Studies In Computer Science And Engineering, U.B.D.T. College Of Engineering, Davanagere.

[7] SPEECH EMOTION RECOGNITION WITH MULTISCALE AREA ATTENTION AND DATA AUGMENTATION Mingke Xu1, Fan Zhang2, Xiaodong Cui3, Wei Zhang3 1Nanjing Tech University, China 2IBM Data And AI, USA 3IBM Research AI, USA

[8] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Towards real-time speech emotion recognition using deep neural networks. In Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on, pages 1–5. IEEE, 2015.

[9] Koteswara Rao Anne, Swarna Kuchibhotla, Acoustic Modeling for Emotion Recognition, Studies in Speech Signal Processing, Natural Language Understanding, and Machine Learning, Springer Briefs in Speech Technology 2015.

[10] M. E. Ayadi, M. S. Kamel, F. Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases”, Pattern Recognition 44, PP.572-587, 2011.

[11] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, “Speech Emotion Recognition Using Hidden Markov Model”, Eurospeech, 2001.

[12] C. M. Lee, S. S. Narayanan, “Towards detecting emotions in spoken dialogs”, IEEE transactions on speech and audio processing, Vol. 13, No. 2, March 2005.

[13] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, “Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,” in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, 2007, vol. 4, pp. IV–957–IV–960.

[14] S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.

[15] B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010.

[16] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Interspeech, vol. 53, pp. 320–323, 2009.

[17] Lin, Y.L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 8, pp. 4898–4901

[18] Nicholson, J.; Takahashi, K.; Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. In Proceedings of the 6th International Conference on Neural Information Processing (ICONIP ’99), Perth, Australia, 16–20 November 1999.

[19] Schüller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004.

[20] Harár, P.; Burget, R.; Kishore Dutta, M. Speech Emotion Recognition with studies. In Proceedings of the 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2–3 February 2017; pp. 137–140.

[21] Anvarjon, T.; Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors 2020, 20, 5212. [CrossRef]

[22] Balti, H.; Elmaghraby, A.S. Emotion analysis from speech using temporal contextual trajectories. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Funchal, Portugal, 23–26 June 2014.

[23] Balti, H.; Elmaghraby, A.S. Speech emotion detection using time dependent self organizing maps. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, Athens, Greece, 12–15 December 2013.

[24] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3):572–587, 2011.

[25] R. Elbarougy and M. Akagi. Cross-lingual speech emotion recognition system based on a three-layer model for human perception. 2013 AsiaPacific Signal and Information Processing Association Annual Summit and Conference, pages 1–10, 2013.

[26] C. Park, D. Lee, and K. Sim. Emotion recognition of speech based on RNN. (November):4–5, 2002.

[27] S. Wu, T. H. Falk, and W. Y. Chan, “Automatic recognition of speech emotion using long-term spectro-temporal features,” in 16th International Conference on Digital Signal Processing, (Santorini- Hellas), pp. 1–6, IEEE, 5-7 July 2009. DOI: 10.1109/ICDSP.2009.5201047.

[28] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion recognition in spontaneous speech using gmms,” in INTERSPEECH 2006 - ICSLP, (Pittsburgh, Pennsylvania), pp. 809–812, 17-19 September 2006.

[29] S. Haq and P.J.B. Jackson. "Speaker-Dependent Audio-Visual Emotion Recognition", In Proc. Int'l Conf. on Auditory-Visual Speech Processing, pages 53-58, 2009.

[30] B. W. Schuller, ‘‘Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,’’ Communication ACM, vol. 61, no. 5, pp. 90–99, 201

[31] Kun Han, Dong Yu, and Ivan Tashev, “Speech emotion recognition using deep neural network and extreme learning machine.,” in Interspeech, 2014, pp. 223–227.