International Journal of Research Publication and Reviews

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

SPEECH EMOTION RECOGNITION SYSTEM FOR ROBOTIC APPLICATION USING MACHINE LEARNING

Husbaan I. Attar[1], Nilesh K. Kadole[1], Omkar G. Karanjekar[1], Devang R. Nagarkar[1], Guide – Prof. Sujeet More[2]

[1,2]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 The Emotions 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.

2. LITERATURE SURVEY

Emotion recognition is a part of speech recognition. There are methods to recognize emotions using machine learning. Emotion recognition is the process of identifying human emotion whereas speech recognition enables the recognition and translation of spoken language into text by computers. This paper explains about machine learning, trends in machine learning, deep learning, and its trends and applications, emotion recognition and speech emotion recognition.

Author	**Year**	**Emotions**	**Type of Corpora**	**Description**
Daniel J.France et al.	2007	Depression and Neutral	Natural	Emotions are recognised from speech to identify depressed and suicide prone patients.
Q.Jin et al.	2015	Anger, Happy, Sad, Neutral	Elicited	Generate features from both acoustic and lexical level to identify emotion like anger, sad, happy, neutral.

Author	**Year**	**Features**	**Description**
Qirong Mao et al.	2014	Pitch, energy	Use convolution neural network to identify emotions like anger, joy, sadness, and Neutral.
Author	Year	Features	Description
Daniel J. France et al.	2000	Pitch, Amplitude Modulation, Formant	Analyse and compare different speeches of male and female patients and diagnose the depression level.
H.K. Palo et al.	2015	MFCC, LPC, LPCC	Analyse two emotion classes like low arousal and high arousal.

Deep Learning:

Deep Learning In A Single Term We Can Understand As Human Nervous System. Machine Vision Deep Learning Sets Are Made To Learn Over A Collection Of Audio/Image Also Known As Training Data, In Order To Rectify A Problem. The Various Deep Learning Models Trains A Computer To Visualize Like A Human.

Deep Learning Models Based On The Inputs To The Nodes Can Visualize. Hence Network Type Is Like That Of A Human Nervous System, With Every Node Performing Under

A Larger Network As A Neuron. So, Deep Learning Models Are Basically A Part Of Artificial Neural Networks. Algorithms Of Deep Learning Learns In Depth About The Input Audio/Image As It Passes Over Every Neural Network Layer. Low-Level Characteristics Like Edges Are Detected By Learning Given To The Initial Layers, And Successive Layers Collaborate Characteristics From Prior Layers In A More Philosophical Representation.

Images, Sounds, Censor Data And Other Data Are Those Digital Forms Patterns Which Deep Learning Recognizes. For Prediction We Are Pre-Training The Data And Constructing A Training Set And Testing Set (Results Are Known). As Our Prediction Obtains An Optimum Node Such That The Predicted Node Provides The Satisfactory Output

Basis Of The Neurons Are In Different Levels And Created To Predict At Every Level And The Most-Optimum Predictions, And Thereafter For The Best-Fit Outcome We Use The Data. It Is Treated As True Machine Intelligence.

A Convolutional Neural Network (CNN) Is A Sort Of Feed-Ahead Artificial Network In Which The Joining Sequence Among Its Nodes Is Motivated By Presenting An Animal Visual-Cortex.

Single Cortical Neurons Give Response To The Stimuli At A Prohibited Area Of Region Known As The Receptive Areas. The Receptive Areas Of Various Nodes Semi-Overlap So That They Can Match The Visual Area. The Reply Of A Single Node For Stimuli Among Its Receptive Area Could Be Mathematically Through The Convolution Operations. Convolutional Network Was Motivated By Natural Procedures And Are Varieties Of Multi-Layer Perceptron Formulated To Use Least Quantity Of Pre-Processing. They Have Broad Use In Image And Video Recognition, Recommendation Systems And NLP.

The Dimensions Of The Characteristics Map (Convolved Features) Is Regulated By Following Parameters:

- Depth: Representing The Filter Count We Used In The Convolution Operation.
- Stride Refers To Size Of The Filter, If The Size Of The Filter Is 5x5 Then Stride Is Equal To 5.
- Zero-Padding: Padding The Input Matrix With 0swas Often Convenient Around The Border, In Order To Apply Filter To 'Input Audio' Matrix's Bordering Elements. Using Zero Padding Size Of The Characteristics Map Can Be Governed

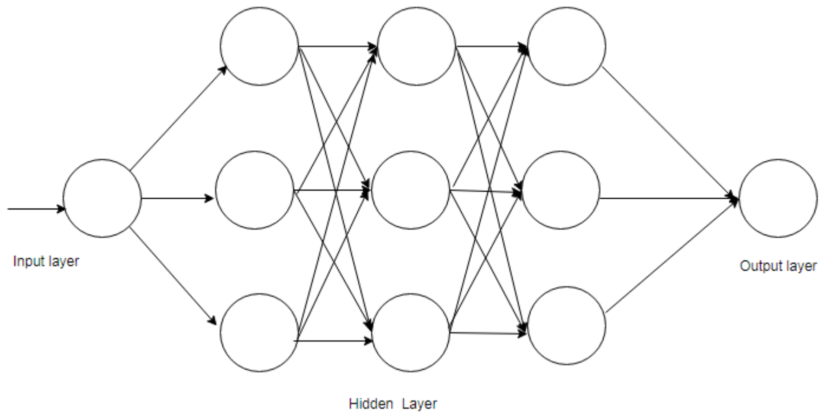

Fig 2.1 Characteristics Of Layers

System Design:

Convolution Neural Networks (CNN's) Are Used To Differentiate The Speech Samples Based On Their Emotion. Databases Such As RAVDEES And SAVEE Are Utilized To Prepare And Assess CNN Models. Kears (Tensorflow's High-Level API For Building And Training Deep Learning Models) Is Used As The Programming Framework To Implement CNN Models. Seven Exploratory Arrangements Of The Present Work Are

Explained In This Section.

Database:

Ryerson Audio-Visual Database Of Emotional Speech And Song (RAVDESS)

The Ryerson Audio-Visual Database Of Emotional Speech And Songs Contains 24 Professional Actors (12 Females, 12 Male), Articulating Two Lexically-Similar Sentence In A Neutral North American Accent. Speech Samples Include Emotions Such As Calm, Happy, Sad, Angry, Fearful, Surprise, And Disgust. Each Emotion Is Produced At Two Stages Of Emotional Intensity (Normal, Strong), With An Additional Neutral Emotion.

Pre-Processing:

The First Step Involves Organizing The Audio Files. The Emotion In An Audio Sample Can Be Determined By The Unique Identifier Of The File Name At The 3rd Position, Which Represents The Type Of Emotion. The Dataset Consists Of Five Different Emotions

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

The Dataset

For This Python Project, We'll Use The RAVDESS Dataset;

This Is The Ryerson Audio-Visual Database Of Emotional Speech And Song Dataset, And Is Free To Download.

Prerequisites

You'll need To Install The Following Libraries With Pip:

Pip Install Librosa Soundfile Numpy Sklearn Pyaudio

3. 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

DATA SET

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

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.

EXISTING SYSTEM

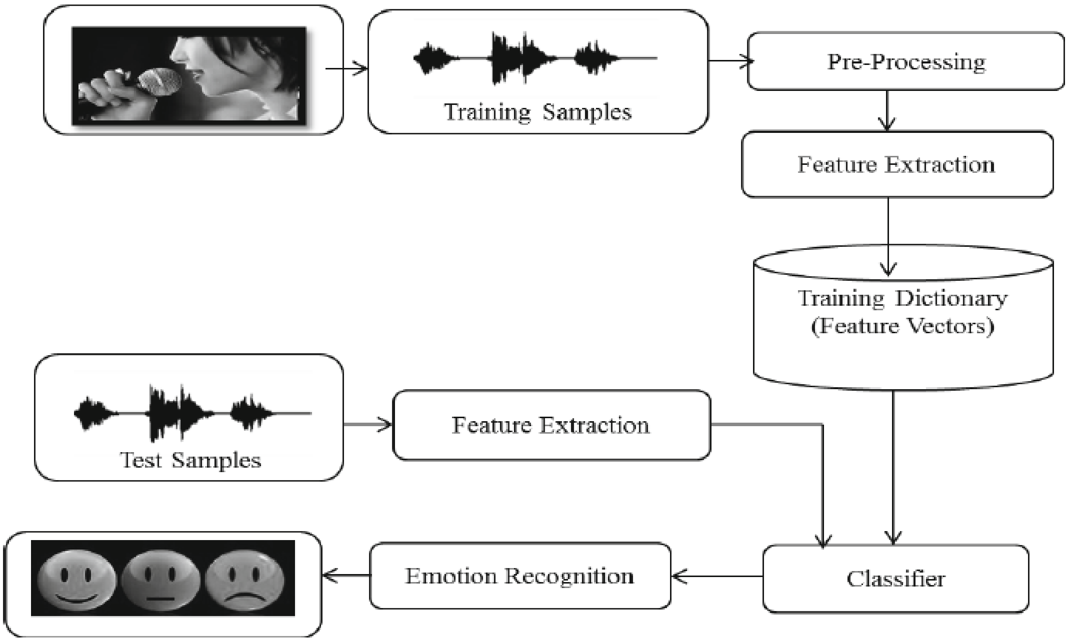

Fig 6.1 Architecture

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

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)

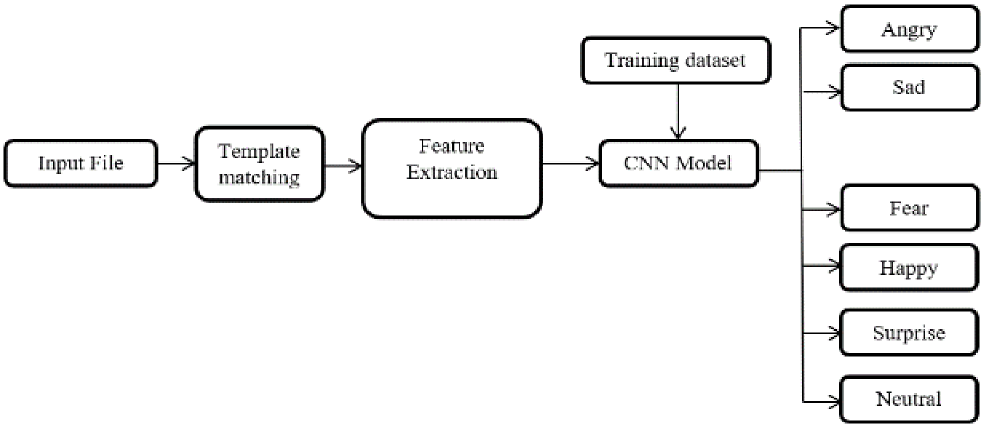

Fig 6.2 CNN Algorithm

4. 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 Abdelwahab And Carlos Busso, Multimodal Signal Processing (MSP) Laboratory, Erik Jonsson School Of Engineering & Computer Science, University Of Texas At Dallas, Richardson, Texas 75083, U.S.A.

[2] Voice Based Emotion Recognition With Convolutional Neural Networks For Companion Robots Eduard FRANT¸ I1, 2, Ioan ISPAS1, Voichita DRAGOMIR3, Monica DASCA˘ 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 And Computer Engineering, Seoul National University, Seoul, Korea Fmysmilesh, Byuns9334, Kjungg@Snu.Ac.Kr

[4] Speech Emotion Recognition Using CNN, Article In International 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 Of Engineering, 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] Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech Emotion Recognition Using Hidden Markov Model", Eurospeech, 2001.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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

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