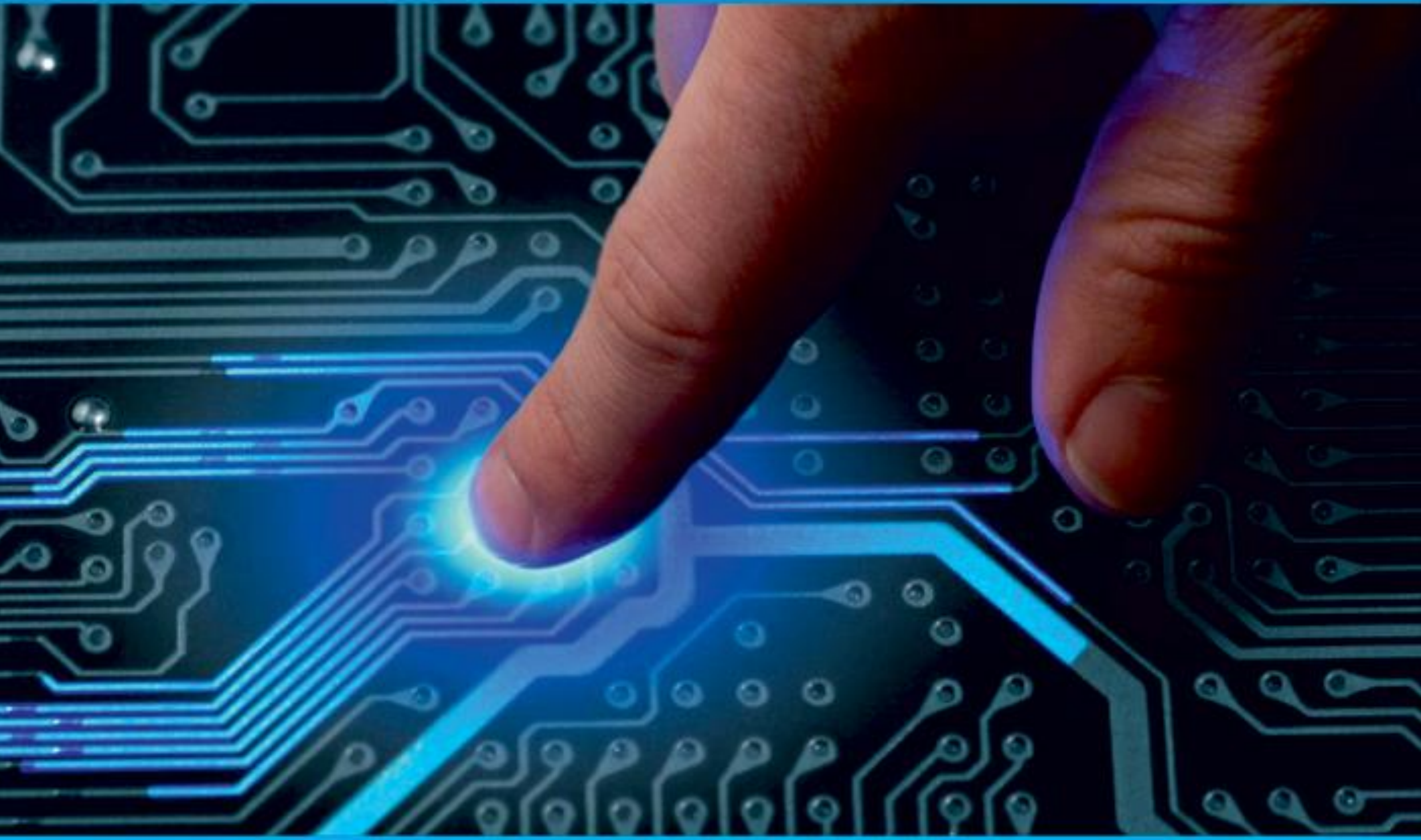
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105240 |

Speech Emotion Recognition Using Machine Learning

Abhishek Poddar, Harsh Raj, Yash Chaudhari, Aditya Mungal, Prof. Kanchan Wankhade

Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Professor, Department of Computer Engineering, SKNCOE, Pune, Maharashtra, India

ABSTRACT: Speech emotion recognition has become a hot topic in the field of human-computer interaction. In order to improve the accuracy of emotion recognition, this paper proposes a new speech emotion recognition technology based on the combination of deep and shallow neural networks. First, the speech signal is preprocessed, then the parallel training sample set is established, and the Deep Belief Network (DBN) is used to automatically extract and recognize the speech emotion features. Finally, the shallow neural network is used to obtain the final recognition results. In order to evaluate the quality of the new method, we compared three systems to identify five emotions, and the experimental results show that the proposed method can effectively improve the accuracy of emotion recognition

People are exchange information and emotions through speech. Emotion recognition from speech is used in many applications such as education, customer service, speech synthesis, medical analysis and forensics etc.The main aim of speech emotion recognition system is to predict the emotions correctly.

KEYWORDS: Speech emotion, Machine learning, Tensor flow, CNN

I. INTRODUCTION

Speech Emotion Recognition (SER) is the process of extracting the speaker's emotional state from the speech. This is used in various applications like call-center services to know the
response of call attendant to a customer, in vehicles to know the psychological state of the person who is driving in order to avoid accidents, as a diagnosing tool to detect various disorders of patients in medical services, in Etutoring and story-telling applications to adapt according to the mood of the listeners, etc. The most important application of SER is in

Teaching & Learning technologies. There is a rapid development in this field from the past few decades. The development in different technologies made the scientific analysis in education field easy. But whatever the learning and teaching methodology may be, the success in learning process depends mostly on the learners' attitude towards the subject or career. In this context, many educationists or psychologists are being appointed in many educational institutes to motivate the students personally or group wise. But, the problem is to identify the emotions of the student when he/ she are present in the class/ library or in canteen speaking to fellow students. This task can be done using an SER system.

II. RELATED WORK

In [1] Conventional standard emotion recognition systems do not apply well for the analysis of true emotional state, mostly because in standard emotion analysis the emotional state of a person is recognised over the complete utterance considering that emotions are mutually exclusive where as in real time it can also be a combination of emotions with a concealed emotion. [2]. In this paper, a novel feature fusion of Teager Energy Operator (TEO) and Mel Frequency Cepstral Coefficients (MFCC), as Teager-MFCC (T-MFCC) feature extraction technique, is used to recognize the stressed emotions from speech signal. TEO is a non-linear feature of speech, basically designed to identify the stressed emotions. In [3] Speech emotion recognition has become a hot topic in the field of human computer interaction. In order to improve the accuracy of emotion recognition, this paper proposes a new speech emotion recognition technology based on the combination of deep and shallow neural networks. First, the speech signal is pre-processed, then the parallel training sample set is established, and the Deep Belief Network (DBN) is used to automatically extract and recognize the speech emotion features. [4] We study universal compression of sequences generated by monotonic distributions. We show that for a monotonic distribution over an alphabet of size, each probability parameter costs essentially bits, where is the coded sequence length, as long as. [5] For the sake of ameliorating the precision of speech emotion recognition, this paper put forward a new emotion recognition technique based on Deep Learning and Kernel Nonlinear PSVM (Proximal Support Vector Machine) to discern four fundamental human emotion (angry, joy, sadness, surprise). In [6] This paper proposes an ensemble classifier based on decision-fusion of multiple SER (Speech Emotion Recognition) models. The one of the multiple SER models used in this work is a typical categorical learning model for classifying the emotion labels, while the others are A/V (Arousal/Valence) models that recognize multiple A/V states based on the Russell's A/V emotion space.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Convolutional layer: Identifies salient regions at intervals, length utterances that are variable and depicts the feature map sequence.
- 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.

- 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.
- Dense layer

B. *Description of the Proposed Algorithm:*

- Audio Feature Extraction and Visualizations. (module01) Characteristics extraction is required for classification and depiction. The audio signal is a 3D signal in which 3 axes indicate time, amplitude and frequency. We will use librosa to analyze and extract

characteristics of any audio signal. (.load) function pulls an audio file and decrypts it into a 1D array which is of time series x, and SR is actually sampling rate of x. By default, SR is 22 kHz. Here I will show one audio file display with the use of (IPython.display) function.

Librosa.display is important to represent the audio files in various forms i.e., wave plot, spectrogram and colormap. Wave plots use loudness of the audio at a particular time. Spectrogram displays various frequencies for a particular time with its amplitude.

- To train the model for accuracy calculation. (module02) Within this module we train the model for accuracy estimations. 1 st, import necessary modules. Then pull the dataset. We will receive the sampling rate value with librosa packages and mfcc function. Thereafter this value holds other variables. Now audio files and mfcc value hold a variable consequently it will add a list. Then zip the list and hold two variables x & y. Then we have represented (x, y) shape values with the use of numpy package. International Journal of Psychosocial 14 Mar 2020 2414
- Implementation process of CNN model. (module03) Speech represented in the form of image with 3 layers. While using CNN, do consider, 1st and 2nd derivatives of speech image with time and frequency. CNN can predict, analyze the speech data, CNN can learn from speeches and identify words or utterances.
- Classification of speech emotions. (module04) When testing we provide the audio input. Next, we run the audio in order to hear with ipython.disply packages. Thereafter plot the audio features with librosa.display.waveplot packages. Extract the Characteristics using

librosa.load. It converts one data frame and display structured form. Further it compares loaded model by predict function batch size 32. Ultimately it displays the output from the audio file what sort of expression/emotion that audio file has.

IV. RESULTS

TABLE 1. TEST CASE 1

Test case ID	Test Case	Test Case I/p	Actual Result	Expected Result	Test Case Criteria(p/f)
001	Model- Enter the wrong username or password click on submit	Username or password	Error comes	Error should come	P
002	Enter the correct username and password click on submit button	Username and password	Accept	Accept	p

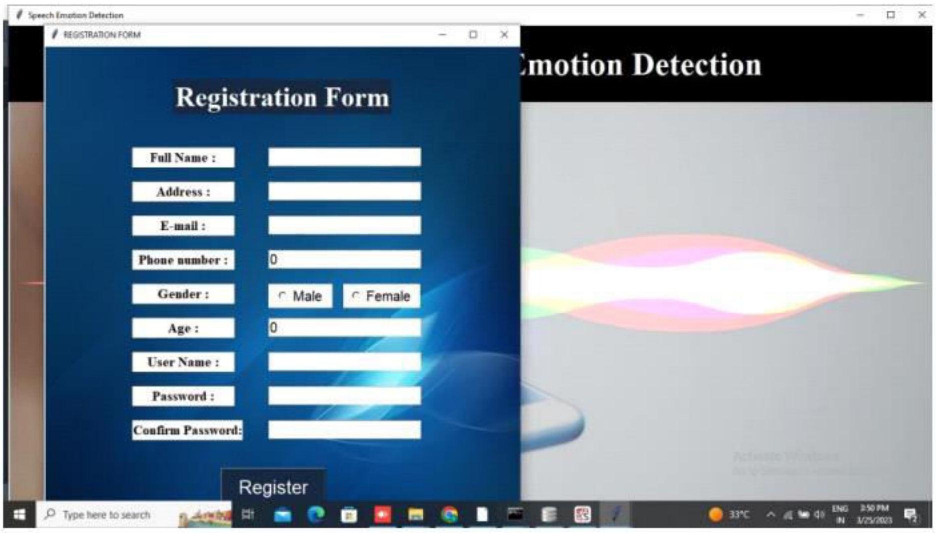
TABLE 2. TEST CASE 2

Test case ID	Test Case	Test Case I/p	Actual Result	Expected Result	Test Case Criteria(p/f)
001	Enter the number in username, Middle name, Last name	Number	Error Comes	Error Should comes	p
001	Enter the character in username, Middle name, Last name field	Character	Accept	Accept	p
002	Enter the invalid email id format in email field	Kkgmail.com	Error comes	Error should come	p
002	Enter the valid email id format in email id field	Kk@gmail.com	Accept	Accept	p
003	Enter the invalid digit no. in phone no. field	99999	Error comes	Error should come	p
003	Enter the 10-digit no. in phone no. field	1234567890	Accept	Accept	p

TABLE 3. TEST CASE 3

Test case ID	Test Case	Test Case I/p	Actual Result	Expected Result	Test Case Criteria(p/f)
001	Store CSV file	CSV file	CSV file store	Error should come	p
002	Parse the CSV file for conversion	Parsing	File get Parse	Accept	p
003	Attribute identification	Check invalid Attribute	Identify Attributes	Accepted	p
004	Weight analysis	Check Weight	Analyze Weight of individual Attribute	Accepted	p

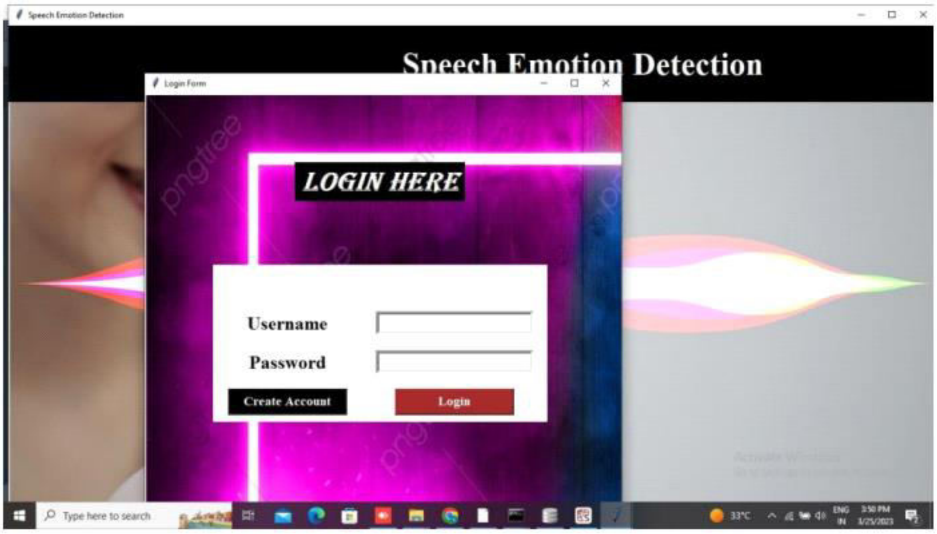
005	Tree Formation	Form them tree	Formation	Accepted	p
006	Cluster Evaluation	Check Evaluation	Should Check Cluster	Accepted	p
007	Algorithm Performance	Check Evaluation	Should Work Algorithm Properly	Accepted	p
008	Query Formation	Check Query Correction	Should Check Query	Accepted	p



Registration Form

Full Name :
Address :
E-mail :
Phone number : 0
Gender : ○ Male ○ Female
Age : 0
User Name :
Password :
Confirm Password:

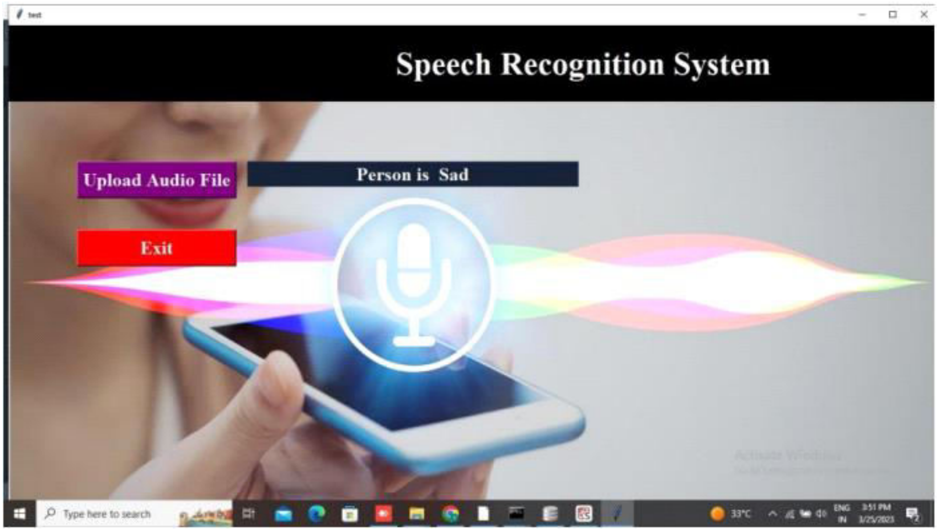
FIG.1. OUTPUT NO 1



LOGIN HERE

Username
Password

FIGURE 2: OUTPUT NO 2



Speech Recognition System

Person is Sad

FIGURE 3: OUTPUT NO 3

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel emotion recognition technique based on deep and shallow neural network to discern five human emotions. We have shown that it is possible to obtain a significant improvement using this method. In particular, it makes good use of the idea of using multiple DBNs to simultaneously automatically extract speech emotion features and identify emotion. But the way of human beings expressing emotions is diverse, it has the expression complexity and culture relative property. There are many limitations for only using speech to recognize emotion. So we can combine facial expression signal to recognize emotion.

REFERENCES

1. R. Cowie , E. Douglas-Cowie, and N. Tsapatsoulis , Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine, 18(1):32–80, 2001.

2.B. Reeves, and C. Nass, The media equation: how people treat computers, television, and new media like real people and places, Cambridge University Press, 1996.

3.L. L. Yu, Z. X. Cai , and M. Y. Chen, Study on emotion feature analysis and recognition in speech signal an overview, Journal of Circuits and Systems, 12(4): 76–84, 2007.

4.J. Nicholson, K. Takahashi, and R. Nakatsu , Emotion recognition in speech using neural networks, Neural Computing and Applications, 9(4): 290–296, 2000.

5.C. H. Park, and K. B. Sim , Emotion recognition and acoustic analysis from speech signal, International Joint Conference on Neural Networks, 2003: 2594–2598.

6. V. Hozjan , and Z. Kacic, Context-independent multilingual emotion recognition from speech signals, International Journal of Speech Technology, 6: 311–320, 2003.

7.W. M. Zheng, M. H. Xin, and X. L. Wang, A novel speech emotion recognition method via incomplete sparse least square regression, IEEE Signal Processing Letters, 21(5): 569–572, 2014.

8.Y. Chen, Z. Lin, X. Zhao, S. Member, G. Wang, and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," pp. 1–14, 2014.

9.L. Chua and T. Roska,"The CNN Paradigm," vol. 4, no. 9208, pp. 147– 156, 1993.

10.X. Xu, J. Deng, E. Coutinho, C. Wu, and L. Zhao, "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition," IEEE, vol. XX, no. XX, pp. 1–13, 2018.

11.z. Huang,J. Epps D. Jaochim ,and V.Sethu "Natural Language Processing Methods for Acoustic and Landmark Event based Features in Speech based Depression Detection", IEEE J.Sel .Top.Signal Process , vol .PP , no. c ,p 2019. . 44

12.J. Deng, X. Xu, Z. Zhang, and S. Member, "Semi-Supervised Autoencoders for Speech Emotion Recognition," vol. XX, no. XX, pp. 1–13, 2017.

13.Y. Qin, S. Member, T. Lee, A. Pak, and H. Kong, "Automatic Assessment of Speech Impairment in Cantonese-speaking People with Aphasia," IEEE J. Sel. Top. Signal Process., vol. PP, no. c, p. 1, 2019. 8. M. D. Zeiler et al., "ON

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105240 |

RECTIFIED LINEAR UNITS FOR SPEECH PROCESSING New York University , USA Google Inc ., USA University of Toronto , Canada," pp. 3–7.

14.C. W. Martin, Ed., The philosophy of deception, 1st ed. Oxford University Press onDemand, 2009, pp. 3– 11, ISBN :9780195327939.

15.P. Ekman and M. O'sullivan, "Who can catch a liar?" in American Psychologist. American Psychological Association, 1991, vol. 46, no. 9, p. 913.

16. S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in 3rd International Conference on Imaging for Crime Detection and Prevention, London, UK, December 3-3, 2009, pp. 1–6.

17.W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: a dataset of spontaneous microexpressions collected from neutralized faces," in 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition(FG), China, April 22-26, 2013, pp. 1–7. 18.P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," inPsychiatry. Taylor & Francis, 1969, vol. 32, no. 1, pp. 88–106.

19.V. Rodellar, D. Palacios, E. Bartolom, and P. Gmez, "Vocal fold stiffness estimates for emotion description in speech," in International Conference on Bio-inspired Systems and Signal Processing, Spain, January 11-14, 2013, pp. 112–119.

20 C. M. Hurley, "Do you see what isee? learning to detect microexpressions ofemotion," in Motivation and Emotion, 2012, vol. 36, no. 3, pp.371–38

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com