iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52752

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Intelligent Speech Emotion Classification Using Deep Learning Technique

Mr. Ashwin Ramteke[1], Harsh Suryawanshi[2], Anuj Zanwar[3], Prathmesh Yadav[4]
[1, 2, 3, 4]*Electronics & Telecommunication Engineering Department, Pune Institute Of Computer Technology, Pune*

Abstract: ***Speech emotion recognition is a very interesting but very challenging human-computer interaction task. In recent years, this topic has attracted a lot of attention. In the field of speech emotion recognition, many techniques have been used to extract emotion from signals, including many well-established speech analysis and classification techniques. In the traditional way of speech emotion recognition, the emotion recognition features are extracted from the speech signals, and then the features, which are collectively known as the selection module, are selected, and then the emotions are recognized, which is a very tedious and time-consuming process, so this paper provides an overview of the deep learning technique, which is based on a simple algorithm based on feature extraction and building a model that recognizes emotions***
Keywords: Speech emotion recognition, deep learning, deep neural network, deep Boltzmann machine, recurrent neural network, deep belief network, convolutional neural network, Long Short Term Memory, Mel-frequency cepstral coefficients.

I. INTRODUCTION

A. Introduction

Speech emotion recognition (SER) is the task of recognizing the emotional aspects of speech independently of the semantic content. While humans can efficiently perform this task as a natural part of voice communication, the ability to use programmable devices to perform this automatically is still a subject of research. Call centre operators and customers, drivers, pilots, and many other users of human-machine communication. gain. Adding emotion to machines is recognized as a key factor in making them look and behave like humans. Robots that are able to understand emotions can display appropriate emotional responses and display emotional personalities. In certain situations, appealing to human emotions can replace humans with computer-generated characters who can have very natural and compelling conversations. Machines need to understand emotions conveyed through language. Without this capability, absolutely meaningful human-machine interaction based on mutual trust and understanding is possible. Machine learning (ML) traditionally involves computing feature parameters from raw data (audio, images, videos, ECG, EEG, etc.). Features are used to train a model that learns to produce the desired output labels. A common problem with this approach is feature selection. In general, the characteristics that most efficiently group data into different categories (or classes) are unknown. Some insights can be gained by testing a large number of different features, combining different features into a common feature vector, and applying different feature selection techniques. The quality of the resulting hand-crafted features can have a significant impact on classification performance. The advent of deep neural network (DNN) classifiers has provided an elegant solution to circumvent the problem of optimal feature selection. The idea is to use an end-to-end network that takes raw data as input and produces class labels as output. There is no need to compute hand-crafted features or determine optimal parameters from a classification standpoint. The network itself does everything. In other words, the network parameters (i.e. the weights and bias values assigned to the network nodes) are optimized during the training process and act as a function to efficiently split the data into desired categories. make. This very useful solution requires a larger number of tagged data samples than traditional classification methods.

B. Scope and Objectives

In this article, we explored how speech data can be used in real-world applications such as automatic speech recognition (ASR) and speech recognition (SER). I researched open source Python packages that support ASR and pitched project ideas. We also considered building a robust SER model using the TORONTO EMOTION SPEECH SET (TESS) dataset to train the LSTM model. This hands-on experience will allow you to start building projects and master SER concepts. Researchers use a variety of speech processing techniques to capture this hidden layer of information so that they can enhance and extract timbral and acoustic features from speech. Converting audio signals to digital or vector format is not as easy as it is for images. The 11 conversion method determines how much key information is retained when exiting the "audio" format. If the given data transformations do not capture softness and serenity, it is difficult for the model to learn emotions and classify patterns.

©IJRASET: All Rights are Reserved

A method for converting audio data into numbers is the mel spectrogram, which visualizes an audio signal based on its frequency content. This is represented as a sound wave and is input to train a CNN as an image classifier. This can be captured by the Mel-Frequency Cepstral Coefficients (MFCC). Each of these data formats has advantages and disadvantages depending on the application. It retrieves data from MFCC and tries to display the data in the appropriate field format used in the model. For example, here we use an LSTM model for feature detection. It takes numbers as input from an MFCC-LSTM model and tries to recognize emotions.

II. DESIGN FLOW/PROCESS

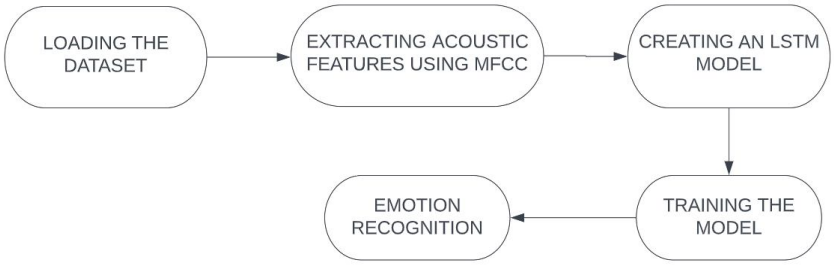


Fig 1: Flow of model Design

1) A set of 200 target words spoken by two actresses (ages 26 and 64) in their career phrase "say the word", each representing seven emotions (anger, disgust, fear, happiness) was recorded. , happy surprise, sadness, neutral) There are a total of 2800 data points (audio files). It contains all 200 target word audio files. The audio file format is WAV format.
2) Then augment the original dataset by creating new data points using existing data and artificially increasing the amount of data.
3) Next, develop a model that will be trained on the collected data. This model is basically a convolutional neural network where the input audio file is passed through a series of filters.
4) Next, extract the acoustic features from the given input audio via MFCC and get the mathematical parameters. Use this to get a collection of pints to base your classification on.

The Following Emotions can be Classified:

- Angry
- Disgust
- Neutral
- Sad
- Pleasant Surprise
- Happy

A. Traditional Techniques

Emotion recognition systems based on digitized speech is comprised of three fundamental components: signal pre-processing, feature extraction, and classification . Acoustic pre-processing such as denoising, as well as segmentation, is carried out to determine meaningful units of the signal . Feature extraction is utilized to identify the relevant features available in the signal. Lastly, the mapping of extracted feature vectors to relevant emotions is carried out . List of nomenclature used in this review paper. by classifiers. In this section, a detailed discussion of speech signal processing, feature extraction, and classification is provided . Also, the differences between spontaneous and acted speech are discussed due to their relevance to the topic. Figure below depicts a simplified system utilized for speech-based emotion recognition. In the first stage of speech-based signal processing, speech enhancement is carried out where the noisy components are removed. The second stage involves two parts, feature extraction, and feature selection.

The required features are extracted from the pre-processed speech signal and the selection is made from the extracted features. Such feature extraction and selection is usually based on the analysis of speech signals in the time and frequency domains. During the third stage, various classifiers such as GMM and HMM, etc. are utilized for classification of these features. Lastly, based on feature classification different emotions are recognized.

B. Database used

The dataset used in this is TESS(Toronto emotional speech set) which consists of 200 target words uttered by two actresses (ages 26 and 64) in their career phrase "Say the word_", a sentence representing each of seven emotions (anger, disgust, fear, happiness). was recorded. , represents pleasant surprise, sadness, or neutral). There are 2800 data points (audio files) in total. The dataset is organized to contain each of the two actresses and their emotions in their own folder. It contains all 200 target words of the audio file. The format of the audio file is WAV format.

C. Feature Extraction

For Extracting acoustic features we use MFCC. Mel frequency cepstral coefficients (MFCC) were originally designed to identify monosyllabic words in coherently spoken sentences, but not to identify the speaker. The MFCC calculation is a replication of the human auditory system to artificially implement the working principle of the ear with the assumption that the human ear is a reliable speaker recognizer. MFCC functions are rooted in the recognized contradiction of the critical bandwidths of the human ear, with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to preserve the phonetically vital properties of the speech signal. Speech signals commonly contain tones of different frequencies, each tone with a real frequency f (Hz), and the subjective pitch is calculated on the Mel scale. The Mel-frequency scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. A pitch of 1 kHz and 40 dB above the threshold of perceptual audibility is defined as 1000 mel and is used as a reference point.

D. Creating Model

Long Short Term Memory (LSTM) is an artificial neural network used in artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTMs have a feedback connection. These recurrent neural networks (RNNs) can process entire data sequences (eg voice or video) as well as individual data points (eg images). This property makes LSTM networks ideal for data processing and prediction. For example, LSTMs can be applied to tasks such as unsegmented and connected handwriting recognition, speech recognition, machine translation, voice activity detection, robot control, video games, and medical care.The name LSTM refers to the analogy that standard RNNs have both "long-term memory" and "short-term memory". The weights and biases of connections in the network change once per training episode, similar to how physiological changes in synaptic strength preserve long-term memories. The network's activation patterns change once per time step, similar to how minute-by-minute changes in the brain's electrical arousal patterns preserve short-term memory. The LSTM architecture aims to provide short-term memory for RNNs spanning thousands of time steps, or "long short-term memory".

III. LITERATURE SURVEY

TABLE I
SURVEY PAPERS FOR DIFFERENT MODELS AND TECHNIQUES IMPLEMENTED

S.No.	References	Emotion recognized	Database used	Deep Learning approach used	Contribution towards Emotion recognition and accuracy	Future Direction
1	J. Zhao et. al (2019) [14]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Berlin EmoDB and IEMOCAP	CNN and DBN with four LFLBS and one LSTM	Deep ID and 2D CNN LSTM to achieve 91.6% and 92.9% accuracy	The presented model can be extended to multimodal emotion recognition
2	E. Lakomkin et. al (2018) [15]	Anger, Happiness, Neutral and Sadness	IEMOCAP database	RNN and CNN	Combine RNN-CNN for iCLub robot and in-domain data with 83.2% accuracy	Future work lead to use may of generative models for real-time data input
3	S. Sahu et. al (2018) [16]	Anger, Happiness, Neutral and Sadness	IEMOCAP database	Adversarial auto-encoders (AAE)	Accuracy with UAR of approximately 57.88%	Future work may lead to recognizing other emotions
4	M. Chen et. al (2018) [17]	Anger, Sadness, Happy, Neutral, Fear, Disgust and Bored	IEMOCAP And Emo-DB databases	3-D CNN LSTM to discriminative features	The model achieves an overall of 86.99%	Future work includes the testing on different databases

5	S. E. Eskimez et. al (2018) [18]	Anger, Frustration, Neutral, and Sadness	USC-IEMOCAP audio-visual dataset	CNN with VAE, AAE and AVB	Better achievement on F-1 score level as 47%	Future work may include RNN
6	W. Zhang et. (2017) [19]	Fear, Anger, Neutral, Joy, Surprise and Sadness	CAS Emotional speech database	Features fusion with SVM and Deep Belief Network for SER is carried out	DBM provides accuracy of 94.6% as compared to SVM that is 84.54%	Future direction may leads to more train DBN with combination of lexical features and audio features
7	P. Tzirakis et. al(2017) [20]	Anger, Happiness, Sadness, Neutral	Spontanous emotional RECOLA and AVEC 2016 Database	Convolutional Neural Network (CNN) and ResNet of 50 layers for both audio visual modality along with LSTM	End-to-end methodology is used to recognize various emotions with 78.7% ac- curacy	Future work involves the application of proposed model on other databases
8	H.M. Fayek et. al (2017) [21]	Anger, Happy, Neutral, Sad and Silence	IEMOCAP Database	Feed forward in combination with Recurrent Neural Network (RNN) and CNN to recognize emotion from speech	Proposed SER technique relies on minimal speech processing and frame based end-to-end deep learning 64.78% accuracy	Future work be to apply the may same model to other databases
9	Y. Zhang et. (2017) [22]	Happiness, Neutral, Disgust, Sadness, Surprise, Fear, and Anger	ComParE and feature eGeMAPS set	multi-tasking DNNs shared with few(MT- SHL-DNN) hidden layers	multi-tasking DNN has been used for the very first time with 93.6% accuracy	Can use multiple Softmax in combination with linear layers for classification and regression tasks
10	Y. Zhang et. (2017) [23]	Neutral, Anger, Fear, Disgust, Sadness, Joy and Happiness	IEMOCAP for emotion recognition and TIMIT for phoneme recognition	Recurrent Convolutional Neural Net- work (RCNN)	This hybrid model has been applied for SER for the very first time with 83.4% accuracy	Possibility of making this model more generic and efficient Cross modal deep learnings
11	Z. Q. Wang and I. Ta- shev (2017) [24]	Happiness, Neutral, Disgust, Sadness, Surprise, Fear, and Anger	Mandarin dataset	DNN based kernel extreme learning ma- chine (ELM)	DNN-ELM approach provides 3.8% weighted accuracy and 2.94% unweighted accuracy in SER	Future works include testing on other corpus

IV. SYSTEM REQUIREMENTS

A. Functional Requirements

1) Dataset: Training Deep learning models requires extensive data to achieve high accuracy, low loss of features and increasing efficiency. The project uses Torronto Emotion Speech Set(TESS).

2) Data Pre-processing: Training Deep Learning models on large datasets requires effective data pre-processing under consideration for effective feature extraction. We have use various data Pre-processing techniques for same.

3) Emotion Recognition Capability: After Pre-processing, a Deep learning model is trained for effective human recognition to generate accurate results.

4) Audio Preprocessing Capability: for real time audio detection, effective audio processing capability is needed to be able to detect and recognize the data conveyed through audio.

5) Model Interpretability: For the effective recognition, effective models like LSTM are used ti explain the working of Models used.

B. Non-Functional Requirements

1) Testability: for effective working of model, modular architecture is used where each part is divided into multiple modules for correct working

2) Reliability: Feature tests are performed to ensure reliability of dataset and quality of dataset, So the accuracy and performance of system are satisfactory.

a) Software Requirements

- Jupyter Notebook
- LSTM Model (Keras Library)

b) Hardware Requirements

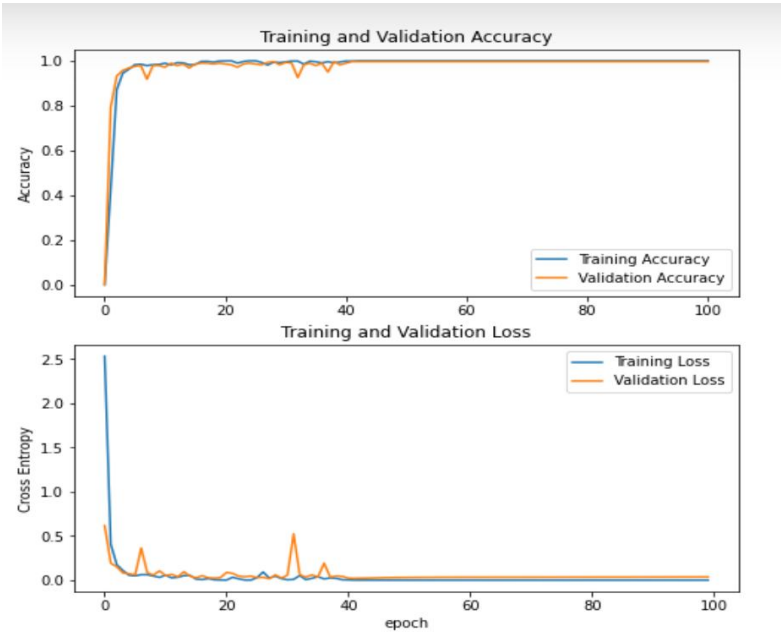
- Processor (i5 or higher)
- GPU: Integrated or Dedicated GPU for processing of dataset and training of model
- RAM : 8GB minimum for processing of dataset and training of mode

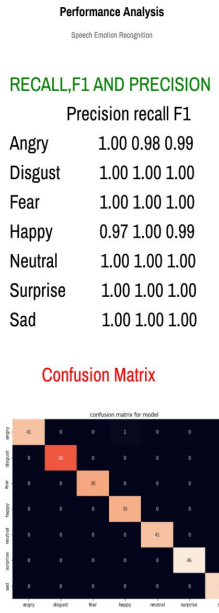
V. RESULTS

The project report presents the results of a speech emotion classification system using Mel-frequency cepstral coefficients (MFCCs) and Long Short-Term Memory (LSTM) networks. The trained model achieved an accuracy of 100% on the test dataset, demonstrating its effectiveness in accurately categorizing emotions from speech signals. The evaluation metrics, including precision, recall, and F1-score, consistently showed high values across different emotion classes. The combination of MFCCs and LSTM networks proved successful in capturing the nuanced variations and patterns associated with different emotional states. The study highlights the potential of this approach in areas such as affective computing and human-computer interaction. Further enhancements and optimizations can improve the system's accuracy and robustness, leading to a deeper understanding of human emotions in real-world scenarios.

Tests were carried out to check the functioning of the following:

- Audio Classifier - Test model with input being an audio from the TESS Dataset.
- Audio Processing - Test processing functionality by processing the audio from the TESS dataset.
- Model Interpretability - Test explain-ability by verifying visualizations generated.





VI. CONCLUSIONS

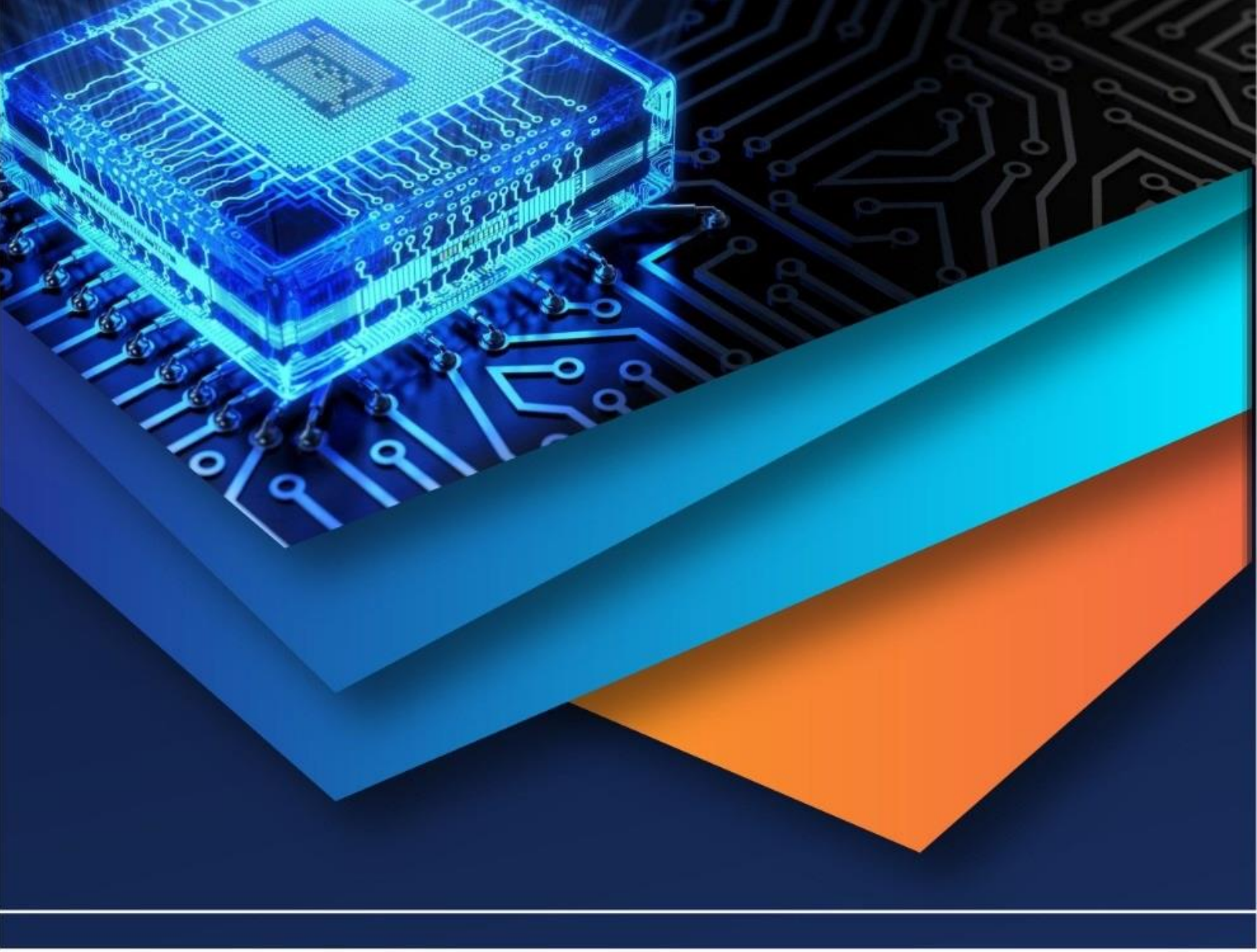
In this project we have tried to analyse some samples of speech using the deep learning technique. Firstly we loaded the datasets then we visualized the different human emotions using our functions wave show and spectrogram using the Librosa library. Then we extracted the acoustic features of all our samples using the MFCC method and arranged the sequential data obtained in the 3D array form as accepted by the LSTM model. Then we build the LSTM model and after training the model we visualized the data into the graphical form using matplotlib library and after some repeated testing using different values the average accuracy of the model is found to be 71%

VII. ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion and it asks would be incomplete without the mention of the people who made it possible and whose constant encouragement and guidance have been a source of inspiration throughout the course of this project. We take this opportunity to express my sincere thanks to my guide respected Mr. Ashwin Ramteke sir for their support and encouragement throughout the completion of this project. Finally, I would like to thank all the teaching and non-teaching faculty members and lab staff of the department of electronics and communication engineering for their encouragement. I also extend our thanks to all those who helped us directly or indirectly in the completion of this project.

REFERENCES

[1] s. Cao et al. [1] proposed a ranking SVM method for synthesize information about emotion recognition to solve the problem of binary classification.
[2] Chen et al. [2] aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method.
[3] Nwe et al. [3] proposed a new system for emotion classification of utterance signals. The system employed a short time log frequency power coefficients (LFPC) and discrete HMM to characterize the speech signals and classifier respectively.
[4] T Wu et al. [4] proposed a new modulation spectral features (MSFs) human speech emotion recognition.
[5] Rong et al. [5] presented an ensemble random forest to trees (ERFTrees) method with a high number of features for emotion recognition without referring any language or linguistic information remains an unclosed problem.
[6] Wu et al. [6] proposed a fusion-based method for speech emotion recognition by employing multiple classifier and acoustic-prosodic (AP) features and semantic labels (SLs).
[7] Narayanan [7] proposed domain-specific emotion recognition by utilizing speech signals from call center application. Detecting negative and non-negative emotion (e.g. anger and happy) are the main focus of this research.
[8] Yang & Lugger [8] presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception from music theory.
[9] Albornoz et al. [9] investigate a new spectral feature in order to determine emotions and to characterize groups.
[10] Lee et al. [10] represent a hierarchical computational structure to identify emotions. Lee et al. [11-12] proposed hierarchical structure for binary decision tree in emotion recognition fields.
[11] Yeh et al. [13] proposed a segment based method for recognition of emotion in Mandarin speech
[12] Dai et al. [14] proposed a computational approach for recognition of emotion and analysis the specifications of emotion in voiced social media such as WeChatt.
[13] El Ayadi et al. [15] proposed a Gaussian mixture vector autoregressive (GMVAR) approach, which is mixture of GMM with vector autoregressive for classification problem of speech emotion recognition.
[14] B. W. Schuller, ‘‘Speech emotion recognition: Two decades in a nut shell, benchmarks, and ongoing trends,’’ Commun. ACM, vol. 61, no. 5, pp. 90–99, 2018.
[15] N. D. Lane and P. Georgiev, ‘‘Can deep learning revolutionize mobile sensing?’’ in Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl., 2015, pp. 117–122.
[16] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, ‘‘Speech recognition using deep neural networks: A systematic review,’’ IEEE Access, vol. 7, pp. 19143–19165, 2019.
[17] K. R. Scherer, ‘‘What are emotions? And how can they be measured?’’ Social Sci. Inf., vol. 44, no. 4, pp. 695–729, 2005
[18] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, ‘‘Emotion recognition in human-computer interaction,’’ IEEE Signal Process. Mag., vol. 18, no. 1, pp. 32–80, Jan. 2001.
[19] R. W. Picard, ‘‘Affective computing,’’ Perceptual Comput. Sect., Media Lab., MIT, Cambridge, MA, USA, Tech. Rep., 1995.
[20] M. El Ayadi, M. S. Kamel, and F. Karray, ‘‘Survey on speech emotion recognition: Features, classification schemes, and databases,’’ Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.
[21] A. D. Dileep and C. C. Sekhar, ‘‘GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,’’ IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421–1432, Aug. 2014
[22] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, ‘‘The automatic recognition of emotions in speech,’’ in Emotion-Oriented Systems. Springer, 2011, pp. 71–99
[23] E. Mower, M. J. Mataric, and S. Narayanan, ‘‘A framework for auto matic human emotion classification using emotion profiles,’’ IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1057–1070, Jul. 2011.
[24] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, ‘‘Prediction-based learning for continuous emotion recognition in speech,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5005–5009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)