Emotion Recognition from Speech Signals Using Machine Learning and Deep Learning Techniques

Gaurav Kataria, Akansh Gupta, V. Sirish Kaushik, and Gopal Chaudhary

1 Introduction

Intelligence is not just about having empirical knowledge, fluid reasoning and high retention power, but also about having the ability to perceive and understand emotions. Hence, to develop a truly intelligent AI system for human-machine interaction, emotional intelligence is a necessity [1]. To understand how to make an intelligent AI, we need to comprehend how humans perceive or recognise emotions. Humans perceive the underlying emotions visibly, audibly and verbally. Similarly, AI systems or emotion recognition systems perceive visible cues such as facial expressions, gestures or body movements from video input, audible cues from speech input and verbal cues in the form of linguistic content. Apart from facial expressions, speech is one of the primary sources for expressing emotions, and thus for a natural human-machine interface, it is important to recognise, interpret and respond to the emotions expressed in this form and will be the primary focus in this paper. The speech input consists of multiple features such as spectral features, prosodic features and linguistic features [2]. Emotion recognition from speech is a challenging task which relies heavily on the combination of speech features to be chosen for extraction and their effectiveness in the classification process. In this paper, both spectral and prosodic features have been utilised.

Nowadays, the ability of AI systems to recognise emotions is a fascinating subject and has wide-ranging applications across many fields, such as clinical studies, call centres, onboard car driving systems, computer games, E-tutoring and virtual assistants. SER systems can be used for audio surveillance of telephonic conversations to help crime investigation. These systems can also be used in E-learning effectively as information about the emotional state of students can help improve

G. Kataria · A. Gupta · V. S. Kaushik · G. Chaudhary (✉)
Bharati Vidyapeeth's College of Engineering, New Delhi, India

S. Srivastava et al. (eds.), *Concepts and Real-Time Applications of Deep Learning*,
EAI/Springer Innovations in Communication and Computing,
https://doi.org/10.1007/978-3-030-76167-7_4

the quality of teaching [2, 3]. This research paper is organised into six sections. Section 1 presents an introduction to the topic of SER systems. Section 2 explores the related work and Sect. 3 explains the methodology of the paper. Section 4 showcases the results achieved by the models and Sect. 5 concludes the research paper. References are mentioned in the last section.

2 Related Work

The problem of choosing the right combination of features to be extracted, database to be used and model architecture to be implemented has been explored by many researchers [4–6], and some of these related to this paper have been presented below.

Lim et al. [7] proposed a novel approach of using a time-distributed convolutional neural network model to recognise emotions from speech. The authors of this paper propose the use of hybrid classifiers as they derive the evidence from different perspectives. A combination of such evidences enhances the performance of the model. The Berlin database was used for testing and training their network. Their proposed approach containing two convolutional layers and two LSTM layers one after the other, employing dropout and other specific hyper-parameters, achieved a precision and F1 score of 88.01% and 86.65%, respectively, on binary classification of emotions. Mirsamadi et al. [8] proposed a novel strategy for feature pooling over time which uses local attention in order to focus on specific regions of a speech signal which are emotionally salient. They used the IEMOCAP dataset and extracted raw spectral features and handcrafted LLDs from 25 msec segments at a rate of 100 frames/sec. They compared the performance of multiple classifiers such as DNN, SVM and RNN with different temporal aggregation techniques. Their proposed strategy outperformed all the other classifiers having a WA of 63.5% and a UA of 58.8%. Sarma et al. [9] proposed the usage of raw speech waveform-based deep neural network for categorical emotion identification. Several different features such as MFCC features, time-domain features and frequency-domain features were fed to a DNN-based model, and the corresponding accuracies were compared. Here, DNN-based model with time-domain features performed the best having a WA of 65.5%. They then trained a TDNN-LSTM-based attention model which achieved a WA of 66.3% and a UA of 60.3% outperforming the rest of the classifiers. In the future, Sarma et al. plan to investigate emotion identification in multidimensional space.

Khanchandani et al. [10] used the Berlin dataset for emotion recognition. Speech samples were noise filtered and then normalised as a part of pre-processing. After this, the entropy and format frequency features were extracted and fed to multilayer perceptron neural network (MLPNN) and generalised feedforward neural network models which achieved accuracies of 89.62% and 98.08%, respectively, on binary classification. Iqbal et al. [11] presented a real-time emotion recognition system, recognising emotions from live recorded speech by analysing the tonal properties. The authors of this paper used RAVDESS and SAVEE datasets and extracted 34

audio features from these samples. They implemented three classifiers, namely SVM, KNN and gradient boosting, where SVM outperformed all the other models.

3 Methodology

The project involves multiple steps, starting with the RAVDESS and SAVEE datasets being imported from [12, 13]. The dataset was pre-processed using standardisation technique, and spectral as well as prosodic features were extracted from the same. Thereafter, the dataset was divided into training and testing sets which were in the ratio of 80% and 20%. Six models were trained and tested on both datasets individually and combined. The classification accuracy of all the six models on each different input variation was noted. The following subheadings further explain the steps in depth (Fig. 1).

3.1 Dataset

This study utilised two datasets, RAVDESS and SAVEE datasets. The RAVDESS dataset contains eight emotions which are calm, happiness, sadness, anger, fearful, surprise, disgust and neutral. These emotions are expressed by 12 male and 12 female actors having a neutral North American accent [12]. The SAVEE dataset also contains the same eight emotions that are portrayed by four English male actors in different intensity variations, i.e. normal and strong intensity [13]. Both of these datasets are actor-simulated emotional speech datasets, and such datasets are quite popular as they have many intrinsic advantages over other speech data collecting techniques such as elicited and natural data collecting techniques. Advantages of such databases are as follows:

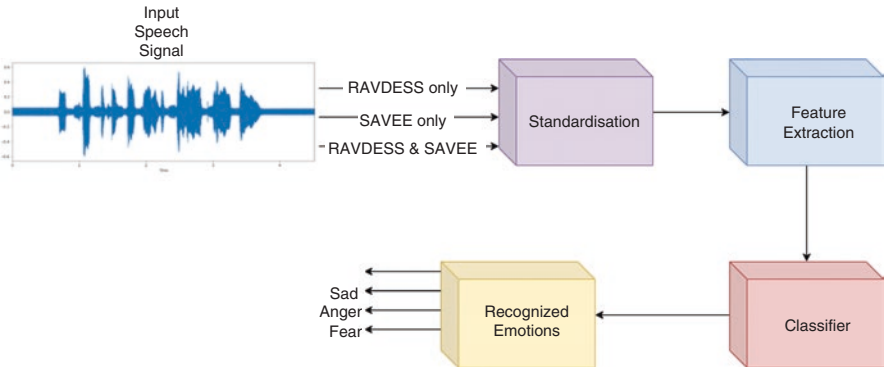

Fig. 1 The block diagram representation of the experiment conducted

(i) As these utterances are recorded in studios, audio signals are of high audio quality and contain no or low amount of noise, which helps the pre-processing stage.
(ii) The emotion expression is balanced out throughout the utterance, and the high intensity of emotion expression makes the classification process easier.

3.2 *Pre-processing and Feature Extraction*

Standardisation was performed on both speech signal databases before the feature extraction process. Data standardisation is the process of rescaling one or more attributes or features so that they'll have the properties of Gaussian distribution with a mean value of 0 and a standard deviation of 1. This is done to make the input data more manageable and preserve its significance. After pre-processing, in the feature extraction process, spectral features and prosodic features were extracted. For spectral features, 40 MFCC, 12 chroma features, spectral centroid, etc. were extracted, whereas for prosodic features, zero-crossing rate, root mean square, etc. were considered [14]. The MFCC features collectively make up the Mel-frequency cepstrum which is a representation of the short-term power spectrum of the input signal. The chroma features represent the tonal content of an audio signal in a condensed form (Figs. 2 and 3).

Root mean square (RMS), zero-crossing rate (ZCR) and spectral centroid (centroid) are defined as:

$$\mathrm{RMS}\left\{x[n]\right\} = \sqrt{\frac{1}{N}\sum_{k} x^2[n]} \tag{1}$$

$$\mathrm{zcr} = \frac{1}{T-1}\sum_{T-1}^{t=1} 1_{\mathbb{R}_{<0}}\left(s_t s_{t-1}\right) \tag{2}$$

$$\mathrm{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{3}$$

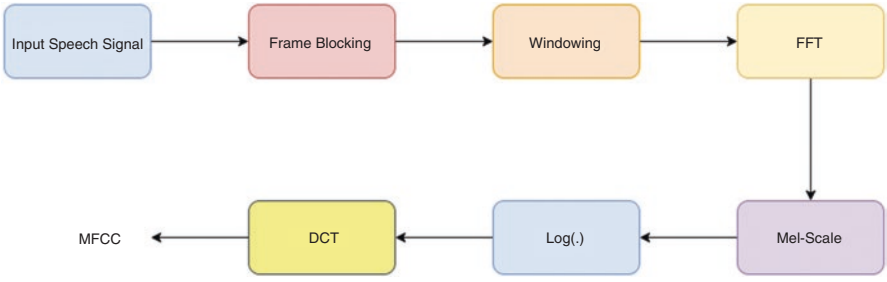

Fig. 2 The procedure to obtain MFCC features

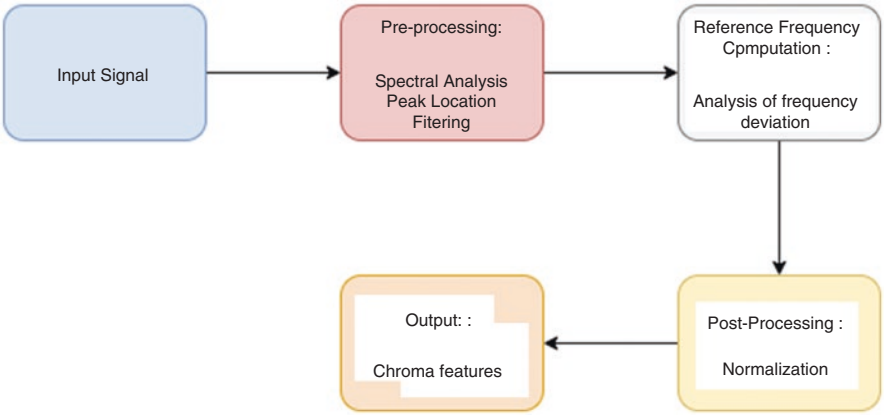

Fig. 3 The block diagram of the procedure of chroma feature extraction

3.3 SVM

Support vector machine (SVM) algorithm is most effective for classification problems. It can be used for linear as well as non-linear classifications depending on the different kernel functions. SVM is an alternative to Bayesian learning. It depends on the support vectors chosen based on their distances. The points which are close to the decision boundary are known as support vectors. The greater the distance of the point from the margin, the higher the confidence. This model completely depends upon the support vectors chosen and the distance metric. In this experiment, SVM algorithm was applied using C = 8 and kernel as 'rbf' parameters. This algorithm outperformed all the other classification algorithms.

3.4 Multilayer Perceptron Neural Network

The multilayer perceptron neural network is a feedforward network and is used for classification problems. This type of neural network has been used widely in speech recognition and image recognition as it has the ability to solve problems stochastically, which often allows approximate solutions for complex problems such as emotion recognition [10]. The MLPNN model presented in this paper consists of 500 hidden layers, with ReLU activation function and constant learning rate of 0.001 and employing Adam optimiser. Batch size used in MLPNN is 16. All the other parameters are set to their default values.

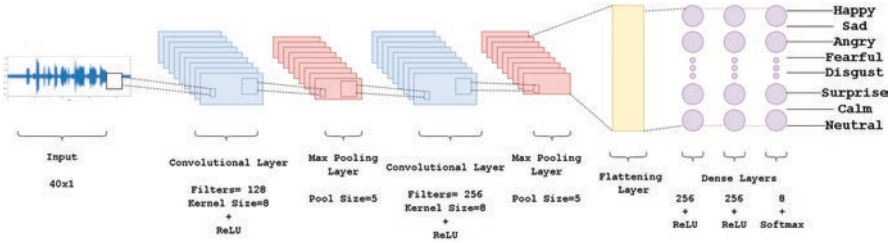

Fig. 4 The architecture of the CNN model

3.5 CNN

The convolutional neural network model presented in this paper consists of two convolution layers with kernel size of 8 and employing ReLU activation function. The first convolutional layer contains 128 filters, whereas the second contains 256 filters. Each of the convolutional layers is followed by a max-pooling layer with a pool size of 5. Followed by these layers are the flattening layer and three dense layers with the first two dense layers having 256 output perceptrons and employing ReLU activation function, whereas the third layer has eight output perceptrons employing softmax function. The following figure shows the architecture of the CNN model (Fig. 4).

The hyper-parameter values used are mentioned in the following table (Table 1).

4 Result

Comparative analysis of performance of all six models based on the classification accuracy achieved while training and testing on the RAVDESS dataset has been presented in Table 2. The CNN model performed the best on the RAVDESS dataset achieving a classification accuracy of 70.83%.

Comparative analysis of all the models for SAVEE dataset is given in Table 3. For this dataset, multilayer perceptron neural network (MLPNN) performed the best, having a classification accuracy of 75% outperforming SVM and logistic regression approaches.

The classification accuracy achieved by the models on the combination of RAVDESS and SAVEE is mentioned in Table 4. Here, SVM and MLPNN achieved the same classification accuracy of 68.22% (Table 5).

ROC curve graph of CNN, SVM and logistic regression models is given below (Figs. 5, 6, and 7).

The conclusion that can be drawn from the above comparative analysis is that MLPNN model performed optimally and consistently, achieving an average accuracy of 70.65% on all three different input variations.

Table 1 The values of hyper-parameters used in the CNN model

Hyper-parameter	Value
Loss function	Sparse categorical cross entropy
Learning rate	0.00002
Dropout factor	0.3
Decay	0.9

Table 2 The accuracy achieved by the models on RAVDESS dataset

Classifiers	Classification accuracy (RAVDESS)
Logistic regression	52.78%
Random forest	61.81%
MLPNN	68.75%
CNN	70.83%
SVM	70.13%
KNN	61.11%

Table 3 The accuracy achieved by the models on SAVEE dataset

Classifiers	Classification accuracy (SAVEE)
Logistic regression	70.83%
Random forest	68.75%
MLPNN	75.00%
CNN	61.46%
SVM	70.83%
KNN	63.54%

Table 4 The accuracy achieved by the models on the combination of RAVDESS and SAVEE datasets

Classifiers	Classification accuracy (RAVDESS+SAVEE)
Logistic regression	44.01%
Random forest	60.68%
MLPNN	68.22%
CNN	65.46%
SVM	68.22%
KNN	58.33%

Table 5 The accuracy achieved by the models on binary classification on the RAVDESS dataset

Classifiers	Binary classification accuracy (RAVDESS)
Logistic regression	93.51%
KNN	93.5%
SVM	96.10%
MLPNN	90.55%

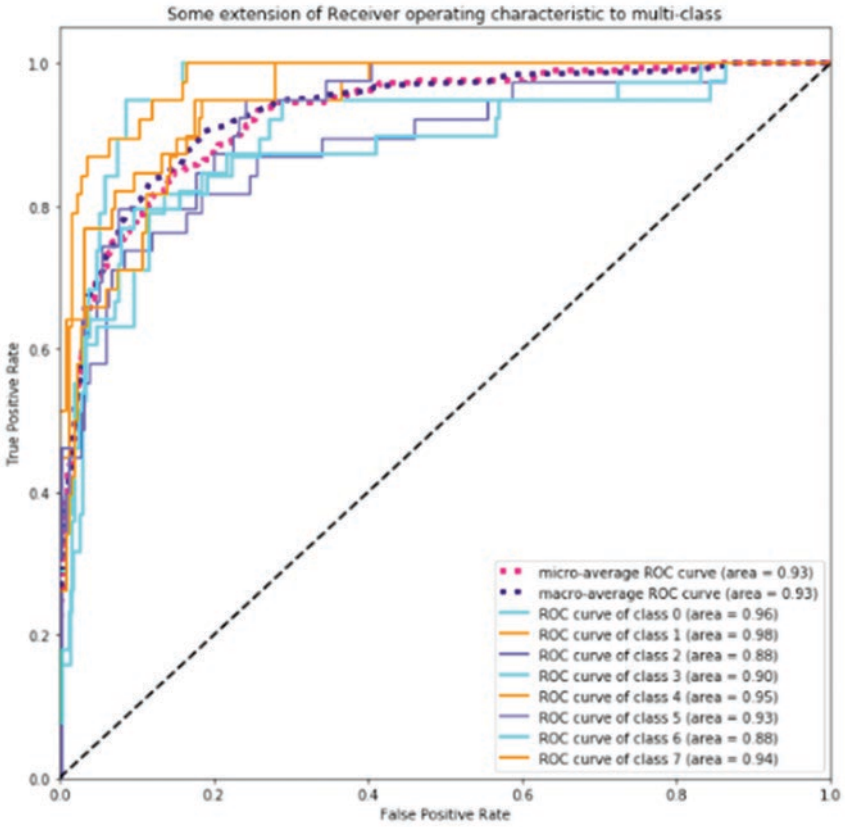

Fig. 5 The ROC curve graph of the CNN model

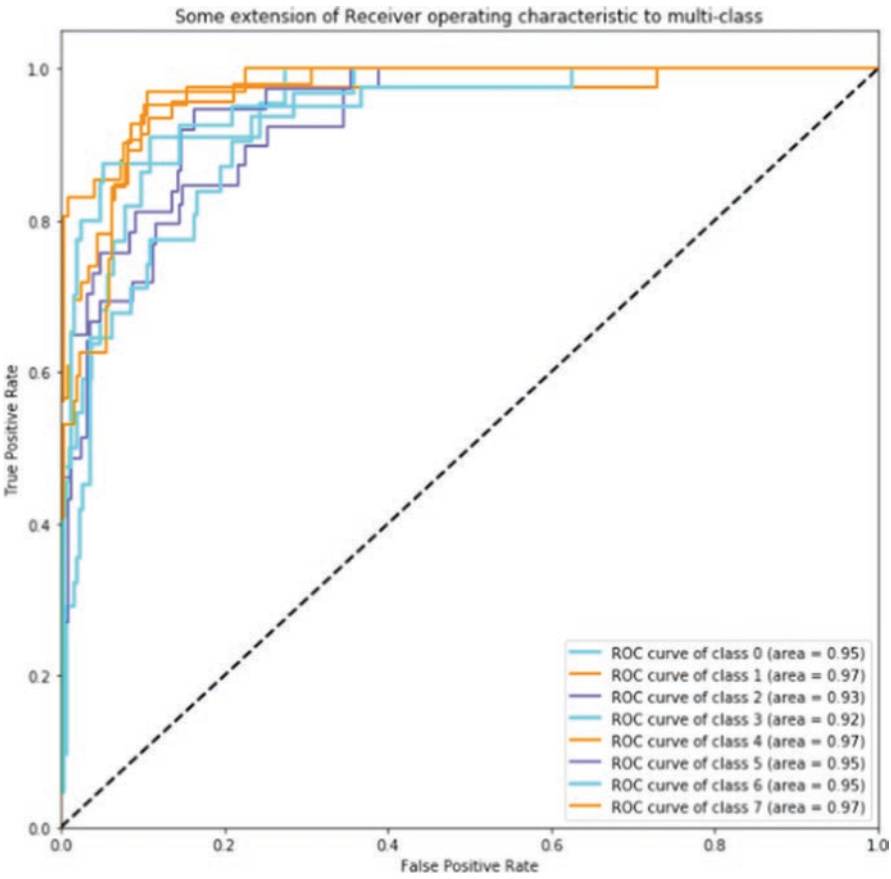

Fig. 6 The ROC curve graph of the SVM model

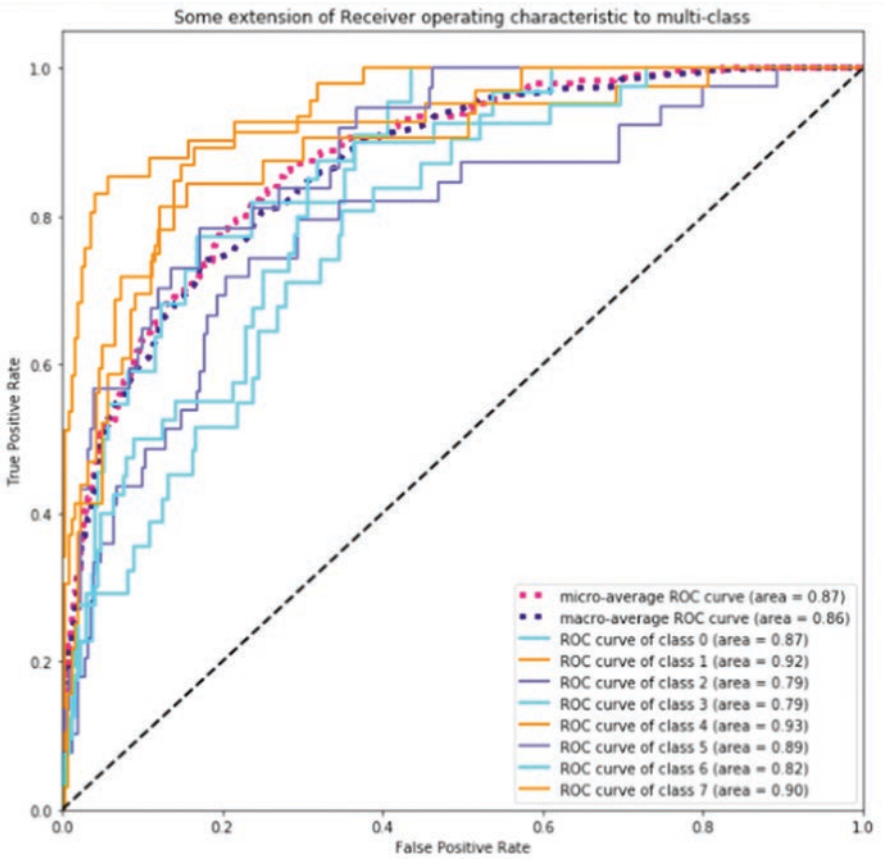

Fig. 7 The ROC curve graph of the logistic regression model

5 Conclusion and Future Work

In this paper, RAVDESS and SAVEE datasets were used to extract spectral and prosodic features. Machine learning and deep learning models were used for classification purposes. Out of the six models, MLPNN performed the best having an average accuracy of 70.65% on three different input variations. Other models such as CNN and SVM also performed optimally achieving a classification accuracy of 70.83% and 70.13, respectively, on RAVDESS dataset.

In the future work, the authors of this paper aim to improve the classification accuracy of all the models by fine-tuning every parameter and hyper-parameter as well as using hybrid classifiers. Overall performance of the models can be improved with the use of larger datasets.

References

1. Huang, K. Y., Wu, C. H., Hong, Q. B., Su, M. H., & Chen, Y. H. (2019, May). Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5866–5870). IEEE.
2. Steidl, S. (2009). *Automatic classification of emotion related user states in spontaneous children's speech* (pp. 1-250). Erlangen, Germany: University of Erlangen-Nuremberg.
3. Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, *15*(2), 99-117.
4. Yacoub, S., Simske, S., Lin, X., & Burns, J. (2003). Recognition of emotions in interactive voice response systems. In *Eighth European conference on speech communication and technology*.
5. Kwon, O. W., Chan, K., Hao, J., & Lee, T. W. (2003). Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*.
6. Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., & Li, X. (2019). Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*.
7. Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1–4). IEEE.
8. Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2227–2231). IEEE.
9. Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2018, September). Emotion Identification from Raw Speech Signals Using DNNs. In *Interspeech* (pp. 3097–3101).
10. Khanchandani, K. B., & Hussain, M. A. (2009). Emotion recognition using multilayer perceptron and generalized feed forward neural network.
11. Iqbal, A., & Barua, K. (2019, February). A Real-time Emotion Recognition from Speech using Gradient Boosting. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1–5). IEEE.
12. http://neuron.arts.ryerson.ca/ravdess/?f=3
13. http://kahlan.eps.surrey.ac.uk/savee/Download.html
14. Padmaja, J. N., & RajeswarRao, R. (2017). Analysis And Identification Of Emotion Specific Features For Speaker Independent Emotion Recognition System Using Gaussian Mixture Models (GMMs). *Advances in Computational Sciences and Technology*, *10*(8), 2491-2505.