Real Time Speech Emotion Recognition using Machine Learning

Nishchay Parikh[1], Khyati Mistry[1], Yashvi Bhavsar[1], AbdulBasit Hakimi[1], Archana Magare[2]

[1]Student, Dept. of Computer Science & Engineering, Institute of Technology & Management Universe, Dhanora Tank Road, Near Jarod, Vadodara - 391510, Gujarat, India
[2]Asst. Professor, Dept. of Computer Science & Engineering, Institute of Technology & Management Universe, Dhanora Tank Road, Near Jarod, Vadodara - 391510, Gujarat, India

---

Abstract -*Emotions play a vital role in day-to-day interactions with other living organisms. Speech Emotion Recognition tries to recognize the emotions from a speech through various techniques and features. This proposed system in the paper can recognize emotions with 78.65% accuracy on RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset with the help of feature extraction techniques that extracts features like MFCC (Mel-frequency Cepstral Coefficients), chroma, and mel spectrogram. The proposed system is also able to recognize emotions in real-time.*

Keywords: *chroma, human speech, mel frequency cepstral coefficients(MFCC), RAVDESS, Multilayer perceptron, Speech emotion recognition*

I. Introduction

Communication plays a vital role in expressing one's notions[1]. Humans interact with each other in various ways such as verbal and non-verbal communication. However, speech is the most effective way to understand what a person wants to convey. Speech is enriched with emotions like happiness, sadness, disgust, anger, and many more depending upon one's mood[1]. Even though technology has developed many individuals are facing problems such as stress and if it continues will result in the worst scenarios of depression and health issues[3]. A solution to this can be by understanding their speech, especially their emotions, as emotions are the carrier of non-verbal communication. As a result, SER comes into the picture and is most famous amongst the researchers[3].

The SER system tries to recognize the emotions and intuitive states from speech. Such systems are advantageous at many places such as call centers, clinics and hospitals, and feedback services from companies[4]. To illustrate, call centers where they listen to customers and try to find out the information from their speech and accordingly improve the services and get happy customers.

SER is a cluster of numerous methodologies that are used in processing and classifying speech signals[5]. Unlike humans who possess Natural Intelligence which includes emotions and consciousness, machines possess Artificial intelligence. The machine needs to learn and compute the and data for this concept like AI and ML are used. Artificial intelligence(AI) is the ability of a computer or a machine(Robot) that are controlled to perform certain tasks like humans and that too with high accuracy, which makes them intelligent machines[6]. Also, Machine learning (ML) is the study of computer programs that can learn, adapt and perform built-in algorithms which allow the software to perform tasks with more accuracy[7]. Moreover, a neural network is a circuit of neutrons and a series of algorithms that recognize deep relationships in a set of data by following processes as a human brain operates.it is a system of artificial neurons[8].

In this paper, the system proposed uses a RAVDESS dataset of 24 actors including both male and female, MLP classifier model for prediction. The accuracy was found to be increased.

2. Literature survey

'

Table 1 - Related Work

Sr no.	Title	Methodology	Findings
1	**Feature extraction algorithms to improve the speech emotion recognition rate-2020[1]**	In this paper, the authors showed various methods but the most efficient one is GMM,which is best for extracting the acoustic characteristics of the speech signal features by ZRC, pitch, and energy.	This paper introduces the significanceof extracting the features using acoustic characteristics like pitch, energy, CWT, and ZCR.

2	**A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition-2019[4]**	In this paper, a new model called Deep Stride CNN architecture that is DSCCN is used. And the test is on the 2 datasets: IEMOCAP and RAVDESS datasets.	It was stated that using the spectrograms on the enhanced speech signals increases the accuracy and decreases the computational complexity. The accuracy of IEMOCAP is increased to 81.75% and dataset RAVDESS is increased to 79.5%. To add, using DSCNN, the size of the dataset is also reduced.
3	**Emotion Recognition of EEG Signals Based on the Ensemble Learning Method: AdaBoost-2021[2]**	In this paper, a method of Emotion recognition using EEG (ElectroEncephaloGram) signals which is based on the ensemble learning method called AdaBoost is proposed. The different domains are considered (like time and time-frequency) and non-linear features related to emotion are extracted from the pre-processed EEG signals and then in the eigenvector matrix, the fused features are given. Also to reduce the dimensions of the feature, a linear discriminant analysis feature selection method is used.	The method proposed in this paper is tested on the DEAP data set and it was proved that this method is effective in recognizing the emotions with the best average accuracy rate up to 88.70%.
4	**Conversational transfer learning for emotion recognition-2020[3]**	In this paper, TL-ERC (Transfer Learning-Emotion Recognition in Conversations) is proposed. Firstly, a generative task of conversation modeling is to be performed using Hierarchical Recurrent Encoder-Decoder (HRED). For the ERC(target) the parameters are then given to classifiers. TL-ERC for ERC used pre-trained affective information from dialogue generators.	In this approach, the datasets used are IEMOCAP and Daily Dialog, and a significant improvement is observed. Also, in the regression task based on the SEMAINE corpus, the improvement is seen. By using the pre-trained weights assists the overall task and produces good and faster Generalisation.
5	**Convolution neural networks for speech emotion recognition-2020[9]**	Dataset - SAVEE(Survey Audio Visual Express Emotions) .Feature Extraction - MFCC(Mel Frequency Cepstral Coefficients) 3. Machine Learning Technique - CNN(Convolution Neural Network)	The accuracy is 84.31% with MFCC+SAVEE better than previous work which shown accuracy between 55%-75%
6	**Speech emotion recognition with deep convolutional neural networks-2020[10]**	Dataset - RAVDESS(Ryerson Audio-Visual Database of Emotional Speech and Song), EMO-DB, IEMOCAP(Interactive Emotional Dyadic Motion Capture) Feature Extraction - MFCC , Chromagram, Mel-scale spectrogram, Tonnetz representation, and spectral contrast features Machine Learning technique - 1-D CNN	The accuracy was found to be 71.61% for RAVDESS, 86.1% for EMO-DB, 64.3% for IEMOCAP thus making the state-of-the-art model outperforms previous works.

7	**Speech emotion recognition using emotion perception spectral feature-2019[11]**	Dataset - CASIA, EMO-DB, FAU AIBO Feature Extracted - custom feature extraction (perception spectral feature) Machine Learning technique - SVM(Support Vector Machine) classifier.	The accuracy was outstanding compared to MFCC feature techniques. Accuracy of 78.6% on CASIA, 81.5% on EMO-DB, and 54.1% on FAU AIBO dataset was to be found which outperforms traditional MFCC feature models.
8	**Speech emotion recognition using Fourier parameters-2015[12]**	Dataset - CASIA, EMO-DB, EESDB Feature Extracted - Fourier parameter features, continuous features, Machine Learning technique - SVM classifier with Gaussian radial basis function kernel	The proposed FP features improved speaker-independent emotion recognition by 16.2 points on the EMO-DB, 6.8 points on the CASIA, and 16.6 points on the EESB database. The performance could be further enhanced by approximately 17.5 points, 10 and 10.5 points by combining the FP and MFCC features on the aforementioned databases hence making a state-of-the-art system.
9	**Cross corpus multilingual speech emotion recognition using ensemble learning-2021[13]**	The model has three layers and the model was also used by authors. One model was trained on multiple languages to check the accuracy of the same model. In the speech emotion database,four different databases have been used. data normalization is also in pre-processing to scale the values. A major analysis was carried out on training & testing of data. a comparative analysis on language was carried out to know language accuracy.	The Urdu database shows increased accuracy by 13%. For EMO-DB(german), the accuracy increased by 8%. For EMOVO (Italian) corpus, the accuracy improved by 11%. Finally, for SAVEE (English) corpus, almost a 5% increase in inaccuracy. All this was obtained using ensemble learning, which uses the most popular three machine learning.
10	**MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach-2020[14]**	To make it more efficient we have used the multi-learning trick approach. Dilated Convolution Neural Network (DCNN) has also been a major part of SER. Two modules RBSC & Seq_L, RBSC for skip connection and the Seq_L for long-term input features. For the final SER to concatenate, a fusion layer was used. IEMOCAP and EMO-DGeneralization datasets are used for obtaining high accuracy. Original speech signals had high accuracy in SER which also used the DCNN and MLT approach.	Here we found that IEMOCAP has 73% accuracy and EMO-DB has 90% accuracy.
11	**Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural network-2021[15]**	Publicly available data-set, CNN SVN . Spiking neural network,Unsupervised learning, Cross data-set evaluation.	Facial Emotion Recognition(FER) has an accuracy of 89% and Speech Emotion Recognition(SER) has 70%.

12	**Augmenting Generative Adversarial Networks for Speech Emotion Recognition-2020[16]**	Generative adversarial networks(GANs), Synthetic and Encoded features, Cross-corpus evaluation method, Synthetic feature vector generation.	Real + Syn accuracy rate, for Sahu et al.[14] we get 45.40 % accuracy, for Bao et al.[16] we get 46.52±0.43 accuracy. Last for Ours we get 46.60 ±0.45 accuracy.
13	**Continuous Speech Emotion Recognition with Convolutional Neural Networks-2020[17]**	In this paper, the authors have used AESDD(Acted Emotional Speech Dynamic Database) to train the datasets with the CNN(Convolutional Neural Network) architecture The CNN input vectors can be 1-dimensional signal as Pulse Code Modulation(PSM) and 2-dimensional signal through spectrotemporal transformations and Mel-scale Coefficients.	The proposed CNN model was more accurate with 69.2% accuracy than the baseline SVM model(60.8%). Data augmentation didn't affect the results a lot, but it showed improved robustness.
14	**Speech Emotion Recognition in Neurological Disorders Using Convolutional Neural Network -2020[18]**	In this paper, a CNN-based model is proposed. The system uses certain tonal properties like RAVDESS and MFCCs, for training the datasets.	The accuracy provided by the proposed model was more than the existing conventional methods and models to incarnate the emotions of a neurologically disabled person. This model helps to create angry, happy, fearful, sad, disgust, neutral and calm emotions into the neurologically disabled.
15	**Speech emotion recognition of Hindi speech using statistical and machine learning Techniques-2020[19]**	In this paper, statistical classification techniques such as Principal Component Analysis(PDA) and Linear Discriminant Analysis(LDA) are used to extract prosodic and acoustic features from speech to recognize emotion.	The local and global features of emotion are extracted for improving the outcomes with the help of PCA and LDA. After applying the KNN and NBC to these features, it can be observed that both techniques provide different outputs but the KNN gives better results than NBC to recognize the emotions- sad, happy and angry.
16	**Speech Emotion Recognition based on SVM and ANN-2018[20]**	In this paper, the features are divided into two parts- acoustic and statistical features which are calculated using the emotional model constructed from SVM(Support Vector Machine) and ANN(Artificial Neural Networks). CASIA Chinese emotional corpus is used to analyze the key technologies of speech and the effect of feature reduction on speech emotion recognition. An algorithm like Principal Component Analysis(PCA) is used to reduce the dimension of features,	The models made on the basis of SVM and ANN are compared with and without PCA. As a result, the SVM model gives more accuracy(46.67% and 76.67& with PCA) than the ANN-based model. Secondly, improvements are observed when doing feature dimension reduction.

3. Methodology

In this proposed system, the first step is to train the model using a RAVDESS dataset and machine learning algorithm; MLP(Multi-Layer Perceptron) classifier[21]. The system works in two modes i.e. takes input from real-time or takes input from the file which is already present.

The proposed system extracts features like Mel-Frequency Cepstral Coefficients (MFCC), chroma, and mel spectrogram from the audio which were used as the input to the MLP classifier model.

The MLP classifier has been used in this proposed system to train and predict emotions like calm, happiness, fearfulness and disgust. By using hyperparameter tuning, the prediction accuracy of the proposed system is increased to its maximum limit.

MLP classifier is a Neural Network used for classification problems. It is a supervised learning algorithm meaning it needs an independent variable(s) and a dependent variable(target variable). Mel-scaled spectrograms and MFCCs are widely utilized in the field of sound classification and speech emotion recognition. These features mimic to a certain extent the reception pattern of sound frequency intrinsic to a human[10]. MFCC is used to convert the conventional frequency of the human speech to Mel Scale. The chroma or chromogram feature is a descriptor , which represents the tonal content of a musical audio signal in a condensed form. The main purpose of chroma features is to aggregate all the spectral information given a pitch into a single coefficient.

Also, the dataset used to train the model is a small part of the RAVDESS dataset which has 24 actors and only a few of the emotions from the entire dataset. However, if we use the whole dataset and the entire range of emotions of the RAVDESS dataset then theoretically the accuracy of the model could be increased.

The accuracy of the aforementioned emotions using the proposed system was found to be 78.65 %. The real-time functionality in the proposed system was achieved using a pyaudio library, which records the audio, and features are extracted and predicted on the same.
The proposed system can also predict the emotions on a recorded audio file and then the feature extraction and prediction is performed.

The proposed system prompts the user to choose from 3 options which are(figure-1) :

1. Create and train model
2. Record and predict emotion
3. Predict emotion on a specified audio file.

When the proposed system is used for the first time then option 1 has to be performed first compulsorily otherwise the prediction cannot be performed on other options.

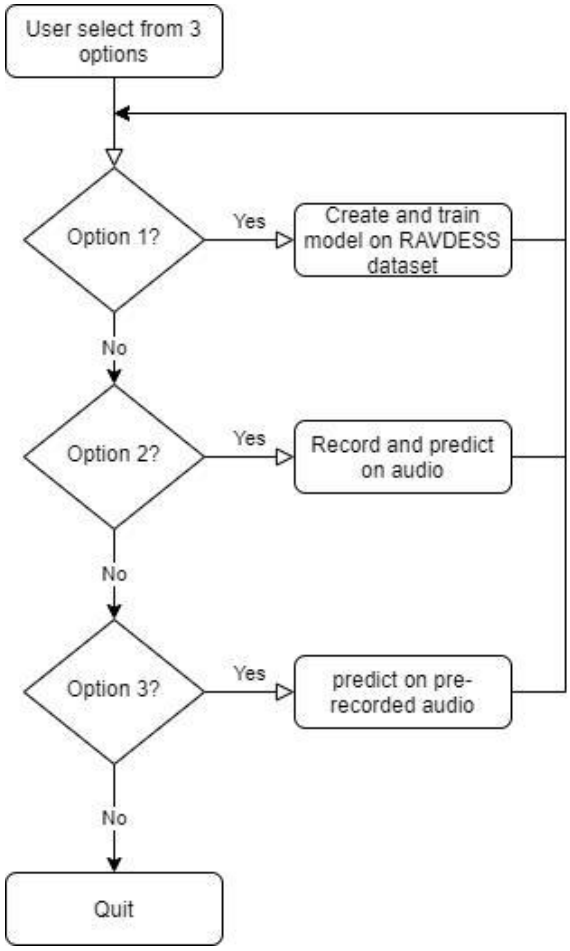

Figure 1 - Concept Diagram-Speech Emotion Recognition

4. Results and Discussion

The proposed system designed a model to train and test the recognition of speech emotion. The table below shows the accuracies of some of the works related to speech emotion recognition mentioned in the literature survey section of this research paper, with the reference number along with proposed work.

Table 2 - Accuracy of Various Algorithms

DATASET / Machine Learning Techniques	ACCURACY
IEMOCAP [10]	73%
CNN SVN [11]	70%
GAN [12]	46.52%
SVM [13]	60.8%
FAU AIBO [7]	54.1%
RAVDESS/LMT[22]	70%
RAVDESS with MLP (Proposed work)	78.65%

With the help of the table, it can be observed that the proposed model, which uses the RAVDESS dataset to train and test the input provided in the form of wav audio files, has increased accuracy.

5. Conclusion

SER is a deeply engaging and trending topic for research in computer science. The proposed system presents a state-of-the-art algorithm for SER with real-time recognition with an accuracy of 78.65%. In the future, the proposed system can be extended to perform emotion recognition on multilingualism. Moreover, it can also be extended to recognize emotions at minute level with the context.

6. References

1. A. Koduru, H.B. Valiveti and A. Budati "Feature extraction algorithms to improve the speech emotion recognition rate" 2020, International Journal of Speech Technology
2. Y. Chen, R. Chang, J. Guo "Emotion Recognition of EEG Signals Based on the Ensemble Learning Method: AdaBoost" 2021, Mathematical Problems in Engineering
3. D. Hazarikaa, S. Poriac, R. Zimmermanna and R. Mihalceab "Conversational Transfer Learning for Emotion Recognition" 2020, Journal of Elsevier
4. Mustaqeem and S. Kwon "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition" 2019, Sensors
5. M. Selvaraj, Dr.R.Bhuvana and S.Padmaja "Human Speech Emotion Recognition" 2016, International Journal of Engineering and Technology
6. Artificial Intelligence - https://languages.oup.com/google-dictionary-en/
7. Machine Learning - https://languages.oup.com/google-dictionary-en/
8. https://en.wikipedia.org/wiki/Neural_ network
9. S. Garg and G. Kumar "Convolution Neural Network for Speech Emotion Recognition" 2020, International Journal of Creative Research Thoughts
10. D. Issa, M. F. Demirci and A. Yazici "Speech emotion recognition with deep convolutional neural networks" 2020, Biomedical Signal Processing and Control
11. L. Jiang, P. Tan, J. Yang, X. Liu and Chao Wang "Speech emotion recognition using emotion perception spectral feature" 2019, Wiley
12. K. Wang, N. An, B. N. Li, Y. Zhang and L. Li "Speech Emotion RecognitionUsingFourier Parameters"2015, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING
13. W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan and T. R. Gadekallu "Cross corpus multi-lingual speech emotion recognition using ensemble learning " 2021, Complex and Intelligent Systems
13. MustaqeemandS.Kwon "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach" 2021, Expert Systems with Applications
14. E. Mansuori-Benssassi and J. Ye "Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural network" 2021, Soft Computing
15. S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak and B. W. Schuller "Augmenting Generative Adversarial Networks for Speech Emotion Recognition" 2020, arXiv
16. N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas and G. Kalliris "Continuous Speech Emotion Recognition with Convolutional Neural Networks" 2020, Journal of Audio Engineering Society
17. S. N. Zisad, M. S. Hossain and K.Andersson"SpeechEmotion Recognition in Neurological Disorders Using Convolutional Neural Networks" 2020, Brain Informatics 13th International Conference, BI 2020
18. A. Agrawal and A. Jain "Speech emotion recognition of Hindi speech using statistical and machine learning techniques" 2020, Journal of Interdisciplinary Mathematics
19. X. Ke, Y. Zhu, L. Wen, and W. Zhang "Speech Emotion Recognition Based on SVM and ANN" 2018,

International Journal of Machine Learning and Computing
20. https://data-flair.training/blogs/python -mini-project-speech-emotion-recogn ition/
21. A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam and I. Zaman, "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames," 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), 2019, pp. 281-285