How to Cite:
Saw, A. K., Arya, C., Sahu, D., & Shrivas, S. (2022). Speech emotion recognition using machine learning. *International Journal of Health Sciences*, *6*(S1), 14313–14321.
https://doi.org/10.53730/ijhs.v6nS1.8662

Speech emotion recognition using machine learning

Mr. Abhishek Kumar Saw
Assistant Professor, Department of Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, Chhattisgarh, India
Email: abhisheksaw7890@yahoo.com

Chetna Arya
B. Tech (Scholar) Department of Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, Chhattisgarh, India
Email: chetna.arya@ssipmt.com

Mr. Devbrat Sahu
Assistant Professor, Department of Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, Chhattisgarh, India
Email: devbratsahu89@gmail.com

Shweta Shrivas
B. Tech (Scholar) Department of Computer Science & Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, Chhattisgarh, India
Email: shweta.shrivas@ssipmt.com

***Abstract*---**Humans connect to each other through language. Verbal words play an important role in communication. The project works on determining an emotion behind verbal words. Speech Emotion Recognition is a system where we determine emotions from live audio. People from all around the globe use speech to convey their emotion irrespective of their background. Emotion recognition from human speech is challenging as there are many factors which play important role in formation of an emotion. It is one of the growing fields in interaction of machine and human. Majorly sub-domains of artificial intelligence are used in the task of prediction. Machine learning is used in the project. Machine learning (ML) uses a dataset and algorithm to predict or detect any future possibility. In this project we propose the application of Artificial Neural Network to determine emotion. Artificial Neural Network is based on how

Manuscript submitted: 18 March 2022, Manuscript revised: 9 April 2022, Accepted for publication: 27 May 2022

biological brain work. It has neurons which are connected to each other and are called nodes. The Classifier used in this project is Multilayer perception (MLP), Decision tree classifier, support vector machine (SVM), random forest classifier. Speech Emotion recognition using machine learning have certain steps to attain result. Firstly we need a dataset to train the program. There are numerous dataset as speech emotion recognition is an evolving field. Then we select the acoustic features that must be extracted. MLP classifier on these features are train the program to determine the project. We are determining emotion of live audio. We have use Google speech recognition for live recording. After live recording the program extract feature and classification helps in determining emotion. Speech emotion recognition (SER) has big future in the growing era of technology.

***Keywords*---**Artificial Neural Network (ANN), Machine Learning (ML), Multiplayer Perception (MLP), support vector machine (SVM), dataset, Speech Emotion Recognition (SER), acoustic features.

1. Introduction

Evolution of human race has witnessed many innovation and discoveries. The most prominent development for human was language. Language is primary and frequent way to communicate with other people. Human speech plays a significant role in effective communication. The best way to understand an emotional state of person is by observing their speech. Emotions are subjective to individual and there are numerous emotions that a person can experience. Therefore, emotion recognition is a tough task. Same emotions in different language have wide degree of difference. There is no standard method to measure and classify them. [1] The emotional state of person transcript in human speech have effected varies factors. Some of these factors that plays important role in formation of a speech are pitch, pace, energy, etc.

With growth of human-machine interaction (HMI), researchers are finding ways for effective communication between machine and humans. Emotions can play important role in this interaction. [3] Here rises the need of development of speech emotion recognition. The better understanding of human emotion can help in determining the mental state of the person. It can also be used to analyze the impact of people's speech on a company during tough, good and excellent times. We can also learn to implement usage of emotion recognition in improving company's service, this can be applicable in different departments where communication plays an important factor such as customer care, telecommunication sector, E-learning, online therapy.

2. Literature survey

Speech Emotion Recognition is used to determine underlying emotion of human speech. The impact of speech emotion can be understood by everyone. Even animals are able to understand the primary emotions like fear and anger. This

was mentioned by Blanton.S on his article "The voice and the emotions". Speech has many factors which effect emotion. The first factor which varies according to different emotions was discovered by Fairbanks.G and Pronovost.W. They found the variation in vocal pitch with simulated emotional speech. Soskin, W.F. and Kauffman, P. E. had provided the information through numerous experiment that voice sounds without any words are fully capable to carry emotional state of a person. After a decade, a new method was brought into light which extract prosodic features called as novel patter technique. This technique gave similar result as human and it was introduced by Dellaert.F, Polzin.T ,Waibel.A. Over the years, a speech analyzer was introduced which can determine angst emotional state using pitch or frequency perturbations. It was introduced by Williamson.J.D. To overcome inadequate measures to determine emotion, a new 10-item mood scale was developed by Watson, David.C, Anna.L, Tellegen, Auke. The idea of Human-machine interaction and the scope of computer to detect human sentiment was brought in the book 'affective computing" by Picard.R.W. This idea was again highlighted and emphasized that now is the time for machine or computer to understand human emotion by Marsella.S and Gratch.J in 2014.[9] It was already known that acoustic features were only accountable for emotion recognition. This myth was broken when researcher added that linguistic information also carries human sentiment. To make emotion recognition more effective, acoustic features and language information should be added in the dataset. This ideology was thus proven true by numerous experiments which solidify the observation that speech emotion recognition operation depend on spokesperson and lingo. [11] Another factor responsible for emotion recognition performance is emotional content. This study was done by Munot.R, Nenkova which observed that neutral emotion is recognized easily than emotional speech. Though in natural communication, emotion factor is of lower significance but necessary for precise outcome. [5] The next decade had majorly worked on enhancing classification, method and algorithm. The improvisation of human-machine interaction can be advanced and is important in current times. [2] Pre-processing of the raw data is essential to gain the relevant data required for training. Different models and classifier are used to find the notable features.[4] The common features extracted for speech emotion recognition are intensity, pitch, speech- rate, formants, pace, energy and Mel-frequency cepstral coefficients. There are feature extraction techniques like linear predictive codes (LPC), Mel Frequency cepstrum analysis (MFFC), Relative Spectral (RASTA), Linear Discriminate Analysis (LDA), Perceptual Linear Predictive (PLP), Principal Component Analysis (PCA).The effectiveness of these techniques was studied by D.Gupta, P.Bansal, K.Choudhary [12].The main problem with speech emotion recognition is the limited amount of dataset on which the model are trained and tested. The method of collecting data is important factor considered by researchers. More dataset are required for training and assessment regarding speech recognition.[6] Usually the model work on supervised learning with labeled data, which left a large gap to be filled as many unlabeled data never get utilized. The solution of it was semi-supervised learning system with co-training algorithm. These techniques had improved the average accuracy by more than 7%.[17][19]. There was implementation of unlabeled speech data in speech emotion recognition by M.Neumann and N.T.Vu. [13]. There were even studies that observed relation between speech features and personality traits by A.Guidi, C.Gentili, E.P.Scilingo and N.Vanello .[7] There have many algorithm that were

tested and applied. Some of them are long-short term memory recurrent neural network (LSTM RNN) by A.Stuhlsatz, C.Meyer, F.Eyben, T.Zielke, G.Meier and B.Schuller [14] and another algorithm was conventional neural network by Q.Mao, M.Dong, Z.Huang, Y.Zhan.[15] To reduce manual labeling workload, they introduced idea by merging semi-supervised and active learning including human help. It also help in labeling the data itself according to its confidence level.[16] The most favored approach regarding emotion recognition which is "continuous dimension approach". This approach leads more toward field of affective computing. It was suggested by Gunesa.H and Schullerb.B.[18]. The current progress of emotion recognition has move towards hybrid algorithm which is combining of different classifier to gain accurate result. [10]

3. Methodology

Machine learning is majorly used to teach the program to train and test on a dataset and increase the accuracy of its prediction. Python language is used for machine learning. Python has big collection of libraries which are useful for statistical application of an algorithm. The modules used are numpy, librosa, sklearn, pickle, pyaudio, soundfile and os. Here we will discuss the proposed system in detail and will have a brief look into our system's method.

1. *Term Used*

The "Speech emotion recognition using machine learning" determines the emotion of a live audio. A speaker will give a statement and the system will determine the speech using machine learning. As we are using machine learning, python language is used for this project.

2. *Working*

Speech emotion recognition uses speech features like Pitch, Energy, MFCC, Chroma , tonnetz as Pattern and recognize it using artificial neural network and multilayer perception classifier. The system contains these important modules - dataset, feature extraction, classifier and recognized output. Basically, the system is based on assay of formation of verbal signals, extracting acoustic features which will be helpful in determining emotional state & taking appropriate classifier to identify states of emotion and give recognized output.

2.1. Dataset

The dataset is most important aspect of machine learning. The project will be trained and tested on the basis of dataset. Dataset as the name suggest is a collection of relevant data. Here we have used the dataset – RAVDESS. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is the most commonly used dataset regarding emotion recognition due to its wide data. The dataset contains voices of 24 professional actor's including 12 female, 12 male actors. Speech dataset have seven emotions whereas song dataset have 5 emotions. This is the more widely used dataset when concerning the emotion recognition. But while working on this project, we faced an issue. The dataset was not equally distributed which resulted in prediction of some emotions more frequent than other emotions. All machine learning project work on dataset. If dataset is uniform then result is also correct. To get more accurate result we added more data of emotions which are fewer in number to balance out all emotions.

2.2. Feature Extraction

Feature extraction included these features of the speech such as Chroma, tonnetz, mel-frequency cepstral coefficient (mfcc), melspectogram (mel) and spectral contrast. These features are extracted and then classified.

2.3. Pre-Processing Module

This is the Pre-Processing module, once after getting the input from user,the input speech is pre-processed. The process of pre-processing is shown in below figure.

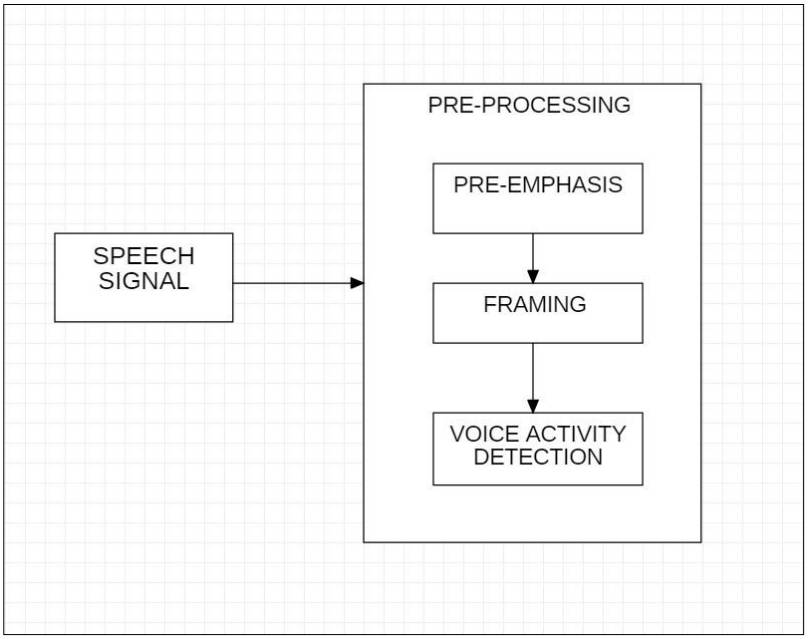

Figure 1- The above figure shows pre-processing module.

2.4. Classification

Multilayer perception(MLP) classifer, Decision tree classifier, support vector machine (SVM), random forest classifier were implemented in Speech emotion recognition. MLP classifier is an artifical neural network which consist of three layer such as input layer,hidden layer and output layer respectively.It is used to detect pattern.The signal flows from input layer to hidden layer and finally to output layer. The MLP classifer is responsible for model training. The testing and training ratio are experimently tested and the best ratio was used with other classifier . The ratio of training and testing can change according to suitablity of model.

3. *Alogrithm*

“Detect emotion algorithm” was used in speech emotion recognition.This algorithm is applied for emotion recognition system. Detect emotion algorithm steps are :-

STEP 1 - Choose the dataset to train and test the data

STEP 2 - Choose which feature should be extracted. Here we have used chroma, tonnetz, mel-frequency cepstral coefficient (mfcc), melspectogram (mel) and

spectral contrast as feature extraction. Librosa and soundfile module are used for the step.
STEP 3 - Set the number of emotions to be detected. We have used primary emotions like sad, happy, angry and neutral.
STEP 4 - Train and test the program on dataset. The ratio of train and test is not fixed. It can be change to best suit the program. Numpy and glob module is used in this step.
STEP 5 - Select the best model parameters determined by grid search. The number of maximum iteration, learning rate and hidden layer size are determined here.
STEP 6 - Initialization of Multilayer perception classifier on basis of model parameter.
STEP 7 - Initialization of Decision tree classifier on basis of model parameter.
STEP 8 - Initialization of support vector machine on basis of model parameter
STEP 9 - Initialization of random forest classifier on basis of model parameter
STEP 10 - Use confusion matrix to predict data and dump model using pickle module.
STEP 11 - We have also added the feature of speech to text using speech recognition module. After training the model, it now ready to use.

4. Result

The project success depends on the level of fineness it have for recognizing the live audio. Dataset plays a prominent part for execution of this project.. Speech emotion recognition was able to understand and detect the underlying sentiment of spokesman. The emotions are categorized in four types – anger, happy, neutral and sad. We have experimented with ratio of training and testing model to get best result. The ratio of 75 % data for training and 25% for testing was observed to give better result than other ratio. We had used grid confusion matrix with MLP classifier, decision tree classifier, support vector machine, random forest classifier. The unweighted accuracy was calculated using average of all emotions accuracy.

Table 1: result of different unweighted accuracy dependent on ratio of training and testing using MLP classifier.

		angry	happy	neutral	sad	Unweighted
Training	testing	accuracy	accuracy	accuracy	accuracy	accuracy
70	30	70.23	69.96	68.43	75.42	71.01
75	25	75.66	72.37	69.54	73.62	72.79
80	20	73.33	63.70	67.83	68.15	68.25

As clearly experimented the training and testing ration 75:25 gives better result. We have used the same ratio to calculate accuracy from support vector machine, decision tree classifier and random forest classifier.

Table 2: it shows unweighted accuracy of emotions from decision tree classifier, support vector machine and random forest classifier.

Emotions	angry	happy	neutral	sad	Unweighted
Algorithms	accuracy	accuracy	accuracy	accuracy	accuracy
Decision tree classifier	73.22	56.67	58.16	71.03	64.77
Support Vector Machine	77.43	73.61	64.89	80.45	74.09
Random forest classifier	75.71	68.07	69.08	68.72	70.39

Table 3: it compares the algorithm used in our project with another algorithm. Data of Personalized Attribute-Aware Attention Network (LC) are taken from PAaAN [19].

Emotions	angry	happy	neutral	sad	Unweighted
Algorithms	accuracy	accuracy	accuracy	accuracy	accuracy
Personalized Attribute-Aware Attention Network (LC)	76.9	71.9	59.1	73.2	70.3
Support Vector Machine	77.43	73.61	64.89	80.45	74.09
MLP classifier	75.66	72.37	69.54	73.62	72.79

5. Conclusion

There is always a possibility of growth. This project is no different, there is possibility of growth where we can use deep learning instead of, machine learning. In machine learning, the weights on neurons are chosen at random each time. Therefore the estimated accuracy varies every time. This situation could be better by implementing deep learning where we can decide the weight on neurons However, executing all the necessary testing, we can say that the project gave satisfactory result. This project was a good experiencing in learning a technology in depth.

6. Futurescope

Global emotion recognition market is predicted to grown in billion dollar industry. Speech emotion recognition can be improved in many ways. We can use different classifier to separate male and female voices for classification to gather more accurate data. The emotion recognition will be useful to track the emotional wellbeing of a person in online counseling. It can also be helpful in E Learning platform. If the teacher is aware about the student state, they can modify the content for better absorption for student. It can also be used to detect whether a person is drunk or not. The addition of emotion detection with facial expression can give more profound result.[8]The development of a classier which can give

accurate prediction in all language will be a great leap towards the era of human-machine interaction.

References

[1] Umair Ayub. “Speech emotion recognition using machine learning”,(2020).

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell benchmarks and ongoing trends", *Commun. ACM,* vol. 61, no. 5, pp. 90-99, Apr. 2018.

[3] M. Chen, P. Zhou and G. Fortino, "Emotion communication system", *IEEE Access,* vol. 5, pp. 326-337, 2017.

[4] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models databases features preprocessing methods supporting modalities and classifiers", *Speech Commun.,* vol. 116, pp. 56-76, Jan. 2020

[5] R. Munot and A. Nenkova, "Emotion impacts speech recognition performance", *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Student Res. Workshop,* 2019.

[6] S. A. A. Qadri, T. S. Gunawan, M. F. Alghifari, H. Mansor, M. Kartiwi and Z. Janin, "A critical insight into multi-languages speech emotion databases", *Bull. Elect. Eng. Inform.,* vol. 8, Dec. 2019.

[7] A. Guidi, C. Gentili, E. P. Scilingo and N. Vanello, "Analysis of speech features and personality traits", *Biomed. Signal Process. Control,* vol. 51, May 2019

[8] F. W. Smith and S. Rossit, "Identifying and detecting facial expressions of emotion in peripheral vision", *PLoS ONE,* vol. 13, no. 5, May 2018

[9] Marsella, S. and Gratch, J. “Computationally modeling human emotion” . Commun. ACM 57, 12 (Dec. 2014).

[10] D. Bharti and P. Kukana, "A hybrid machine learning model for emotion recognition from speech signals", *Proc. Int. Conf. Smart Electron. Commun.,* pp. 491-496, Sep. 2020

[11] Bhaykar, M., Yadav, J. and Rao, K.S. “Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM”. In Proceedings of the National Conference on Communications. (Delhi, India, 2013).

[12] D. Gupta, P. Bansal and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition" in Speech and Language Processing for Human-Machine Communications, Singapore:Springer, 2018.

[13] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2019.

[14] Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G. and Schuller, B. “Deep neural networks for acoustic emotion recognition: Raising the benchmarks” .In Proceedings of ICASSP. (Prague, Czech Republic, 2011). IEEE,5688–5691.

[15] Mao, Q., Dong, M., Huang, Z. and Zhan, Y. “Learning salient features for speech emotion recognition using convolutional neural networks” IEEE Trans. Multimedia 16, 8 (2014).

[16] Leng, Y., Xu, X., and Qi, G. “Combining active learning and semi-supervised learning to construct SVM classifier” Knowledge-Based Systems 44 (2013).

[17] Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. “Semisupervised Autoencoders for Speech Emotion Recognition” IEEE/ACM Transactions on Audio, Speech, and Language Processing 26, 1 (2018).
[18] Gunes, H. and Schuller, B. “Categorical and dimensional affect analysis in continuous input: Current trends and future directions” , Image and Vision Computing 31, 2 (2013),.
[19] Jeng-Lin Li1, Chi-Chun Lee, “Attentive to individual: A multimodal emotion recognition network with personalized attention profile,” Proc. Interspeech, pp. 211–215,(2019).
[20] Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.