Speech Emotion Recognition in Hindi

Name: Tejaswini S
Computer Science and Engineering
RV College of Engineering
Bengaluru
USN:1RV17CS173

Name: Mahathi Siddavatam
Computer Science and Engineering
RV College of Engineering
Bengaluru
USN:1RV17CS082

Name: Khushi Talesra
Computer Science and Engineering
RV college of Engineering
Bengaluru
USN:1RV17CS076

***Abstract*— In today's digital world, understanding emotions play an increasingly vital role in communication, thus making its detection and analysis equally important. The research in the field of emotion recognition is vast and multiple approaches have been undertaken. The focus of our paper lies in the sub-field of emotion recognition from speech signals. In this paper, we have proposed a model to recognize emotion from speech which is in Hindi. The database for the speech emotion recognition system is the IIT (Kharagpur) Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC). Our approach uses the combination of mel frequency cepstral coefficients (MFCCs), chroma and mel spectrogram frequencies, to identify the underlying emotions. The proposed Multi-Layer Perceptron Classifier can classify the emotions - anger, disgust, fear, happy, sad, sarcastic and surprise. The model was able to analyze the tone and pitch of various audio clips of both, male and female voices and the final accuracy obtained with MLP Classifier was 81.52%.**

Keywords—Speech Emotion, Machine Learning, Hindi

I. Introduction

As human beings, we convey emotions through facial expressions, body language, speech and other gestures. These emotions can often vary from person to person, time to time and perception to perception, thus making them highly subjective. Expressing emotions through speech is one of the most common forms of communication. Its popularity is hugely due to its ease and its ability to add a "personal touch". This "personal touch" is what ties to emotion conveyed through speech.

The world is now a global village, and people are more connected to each other than ever before. Technological advancements like the Internet and 4G/5G connections, are major causes that made remote connections more accessible around the globe. These remote communications have emphasized the importance of detection and analysis of emotions, in order to improve the process.

Over the years, there has been a stark increase in human-machine interaction, along with the matching demand of recognizing the underlying emotions. There have been multiple studies that worked on creating an intelligent machine that can recognize human sentiments and simulate appropriate responses or functions based on the same. Cognitive response analysis, robotics, customer relationship systems, music industry, interpersonal relationship analysis and education industry are a few fields where emotion recognition has proved its usefulness.

Deducing emotion through speech can be done in verbal and non-verbal ways - by (i) analysing the semantics of the spoken words and (ii) by analysing the speech signal (pitch, tone, power). To implement method (i), the model must be able to identify sentiment determining keywords from the spoken dialogues and deduce the emotion accordingly. To implement method (ii), the model must be able to inspect various spectral and prosodic features and form a relationship between the acoustic features and the emotions.

The subjective nature of emotions poses a major hurdle in identifying them with certainty. Many factors can influence how emotions are portrayed by a person. Cultural background, mother tongue language, country of residence, gender and disabilities can influence one's delivery style. Delivery style varies in terms of rate of speaking, grammatical structure, voice modulations, pitch and tone. These factors can make recognition very ambiguous. Even human beings sometimes face trouble in understanding another person's emotions. Thus, our machine will require a good amount of learning before it can begin recognizing emotions.

Classification of emotions can be done using classifiers like Gaussian Model, K-Nearest Neighbour Model and Artificial Neural Networks. In this paper, classification is done using a Multi-Layer Perceptron (MLP) Classifier which investigates the acoustic features of the input audio files taken from the IITKGP-SEHSC dataset. It employed a stochastic gradient-based weight optimization function called 'adam' and an activation function for the hidden layers called 'logistic sigmoid function'. The loss curve and accuracy were recorded for the same parameters.

II. Literature Review

In the paper presented by Shashidhar G. Koolagudi, Ramu Reddy, Jainath Yadav, K. Sreenivasa Rao [1] properties of dataset and choice of various classifiers are reviewed. Various prosodic and spectral features of speech are analyzed which are helpful in the further investigation of modern methods of emotion recognition. Various important issues in speech emotion recognition system such as the extraction of most appropriate features from speech and a classifier which recognizes emotions from the speech signal are discussed in [2]. How softness and loudness of the voice can play a very important role in distinguishing the emotions and can impact the accuracy of the model severely has been demonstrated in [9] and finally how various deep learning techniques such as RNN, CNN, DBN etc. impact speech emotion recognition by offering easy model training as well as the efficiency of shared weights was shown in [8]. It formed a base to evaluate the performance and limitations of current deep learning techniques.

III. Methodology

A. *Dataset pre-processing*

i)Audio File Analysis:

The dataset contained 4 folders - 2 of male and 2 of female audio. Each folder was composed of 10 sessions, where each session had 15 audio files for each of the 7 emotions. All audio files were named to suit a supervised learning method. Audio files were visualized to understand how they vary from emotion to emotion.

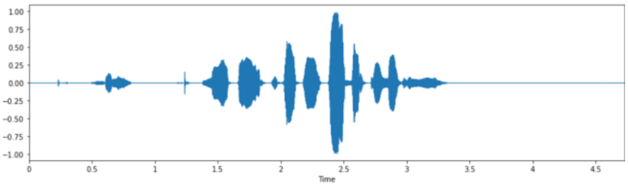


fig 1) Audio waveform of anger

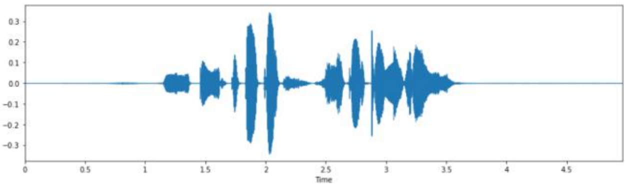


fig 2) Audio waveform of fear

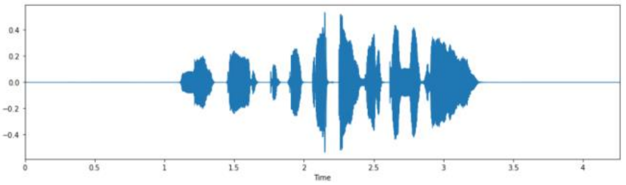


fig 3) Audio waveform of sarcastic

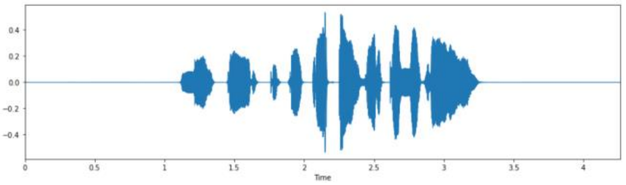


fig 4) Audio waveform of disgust

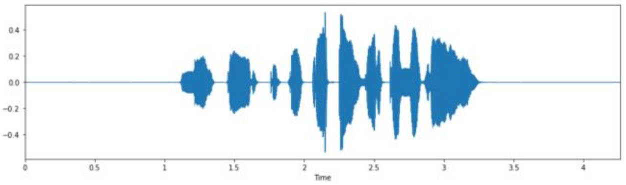


fig 5) Audio waveform of happy

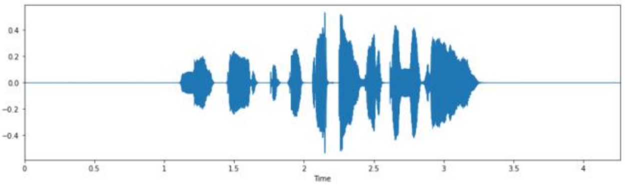


fig 6) Audio waveform of sad

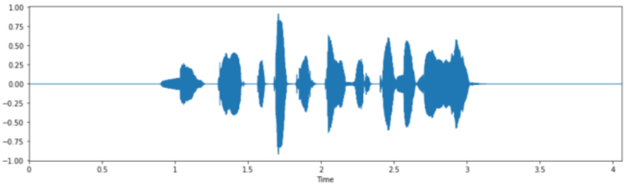


fig 7) Audio waveform of surprise

ii)Feature Extraction and Selection

There are three major types of features:
1)Temporal (time domain): These are simple physical interpretations, that can be analysed as timeframe, like the energy of signal, zero crossing rate, maximum amplitude and minimum energy.
2)Spectral (frequency based): By applying short-term Fourier series to a time-based signal, spectral features like fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off are obtained.
3)Prosodic (suprasegmental based): These are intonations, stress and rhythm that are found at word junctures or over consonants and vowels.
Among these features, our paper focuses on combining the effects of Mel Frequency Cepstral Coefficients (MFCCs), Chroma and Mel spectrogram frequencies. The short-term Fourier series (STFT) of every audio file is extracted. Assuming x : Z → R is a real-valued discrete signal obtained by equidistant sampling with respect to a fixed sampling rate Fs given in Hertz (Hz). Also assume that w : [0 : N − 1] := {0, 1, . . . , N − 1} → R is a discrete-time window of length N ∈ N and let H ∈ N be a hop size parameter [3,4]. Thus, the STFT X of the signal x is (with n ∈ Z and k ∈ [0 : K]) given by

$$\mathcal{X}(m,k) := \sum_{n=0}^{N-1} x(n+mH)w(n)\exp(-2\pi ikn/N)$$

Suitable sample rate is obtained using soundfile library. MFCC is extracted using librosa, which returns 40 values. First, the signal is sliced into short frames and power spectrum of each frame is calculated by applying short-term Fourier transform. Mel filterbank is applied to the power spectrums and the energy in the filter is summed. Then, the discrete cosine transform (DCT) of the log filterbank energies is taken[8]. The formula to convert f hertz into m mels [3,4] is:

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right)$$

Chromagram is computed from the waveform, using the short-term Fourier series obtained earlier. Given a pitch-based log-frequency spectrogram YLF : Z × [0 : 127] → R≥0 as defined in YLF(m, p) := X k∈P (p) |X (m, k)| 2 , a chroma representation or chromagram Z × [0 : 11] → R≥0 can be derived by summing up all pitch coefficients that belong to the same chroma (for c ∈ [0 : 11]) [3,4]:

$$\mathcal{C}(m,c) := \sum_{\{p\in[0:127]\,|\,p \bmod 12 = c\}} \mathcal{Y}_{\mathrm{LF}}(m,p)$$

Mel Spectrogram is also computed using librosa. The mel scale is calculated so that two pairs of frequencies separated by a delta in the mel scale are perceived by humans as being equidistant. The means for mfcc, chroma and mel are individually calculated.

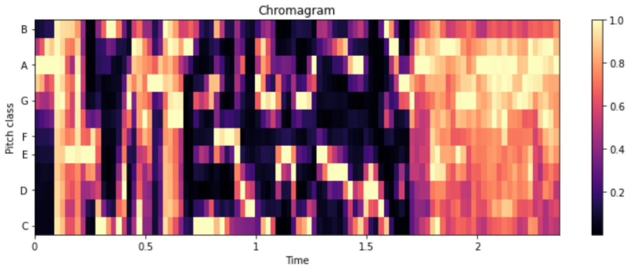


fig 8) Visualization of chroma feature of one of the anger speech audios

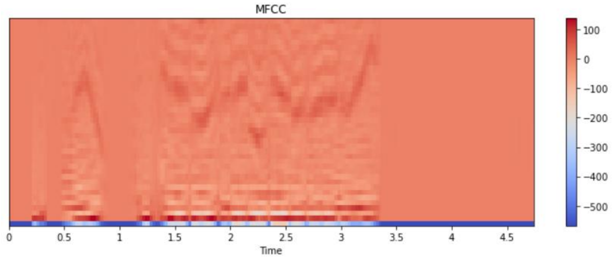


fig 9): Visualization of MFCC feature of one of the anger speech audios

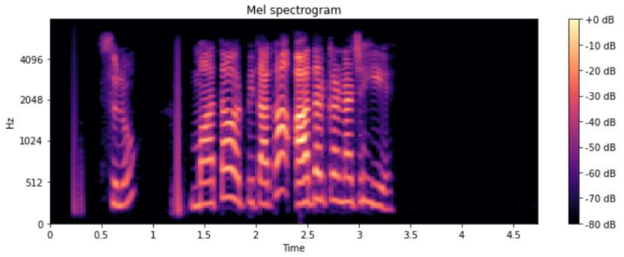


fig 10) Visualization of Melspectrogram feature of one of the anger speech audios

B. *Classifier Selection*

Speech emotion recognition is a simple case of supervised learning in which the classifier must find relationships between the extracted features and the sentiments, thus being able to classify any new and unseen instance fed to it. Different classifiers exist that can be used to perform the task. Gaussian Model, K-Nearest Neighbour Model and Artificial Neural Networks are a few examples. In this paper, we use a class of feedforward artificial neural networks called multilayer perceptron classifier, MLPClassifier in short. Like all neural networks, the MLPClassifier is made up of an input layer, one or more hidden layer and an output layer [6]. Our model is initialized with one hidden layer of size 900. Each node, in the hidden layers and output layer, is a neuron that requires a nonlinear activation function. The activation function used is a logistic sigmoid function, which is a common S-shaped curve (sigmoid curve) with equation [7]:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

where, x_0 is the x value of the sigmoid's midpoint

L is the curve's maximum value

k is the logistic growth rate or steepness of the curve

Weight optimization techniques are algorithms that every neural network uses to decide weights, learning rates and other parameters to ensure reduction in loss. Our MLPClassifier uses 'Adam' as its weight optimization function. Adam optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. Having 3150 instances, our training set was relatively large. Hence, we chose the Adam function as the solver, which converged within the maximum number of iterations of 1000. The L2 penalty parameter 'alpha' was set to 0.01 [5]. An adaptive learning rate was utilized to keep the learning rate constant as long as training loss keeps decreasing [5]. Thus, the MLPClassifier is initialized and can be fitted.

IV. *RESULTS*

The accuracy obtained after the Multi-Layer Perceptron model predicted on the test set was 81.52%. The loss incurred at the end of validation was 0.05015863885328556. The best loss was 0.047398709998271296. The loss curve as seen in Figure 11, was plotted for the performance on the test set.

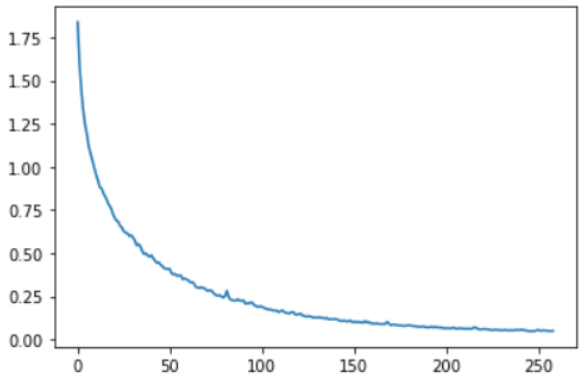
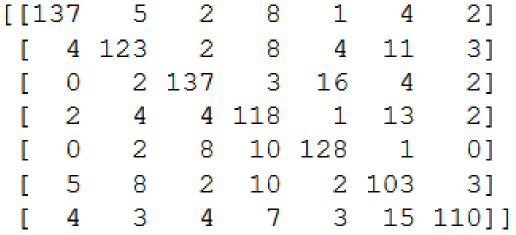


fig 11) Loss curve for test set

The confusion matrix as seen below, was helpful in understanding how the model performed while predicting different emotions. The confusion matrix is also visualized as a heatmap in Figure 12.



```
[[137   5   2   8   1   4   2]
 [  4 123   2   8   4  11   3]
 [  0   2 137   3  16   4   2]
 [  2   4   4 118   1  13   2]
 [  0   2   8  10 128   1   0]
 [  5   8   2  10   2 103   3]
 [  4   3   4   7   3  15 110]]
```

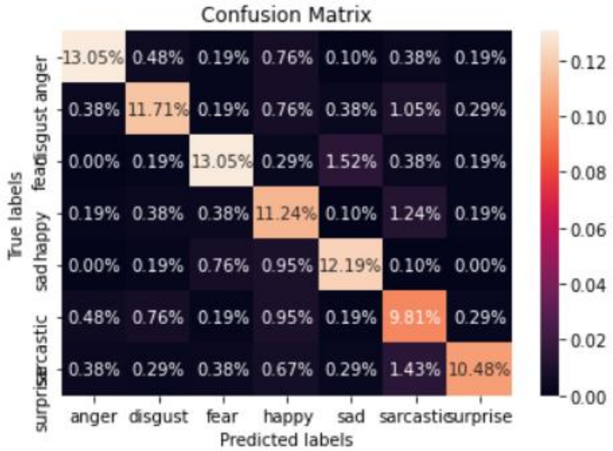
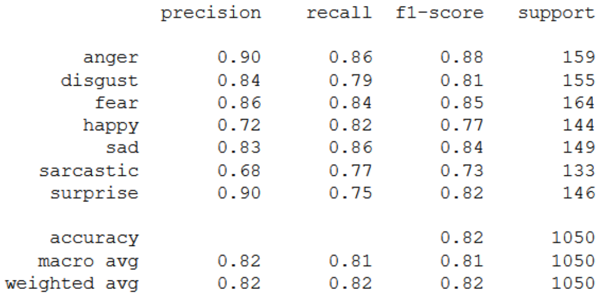


fig 12) Heat map

The classification report as seen in Figure 13, shows the precision, recall, f1-score, support for each emotion. The macro average and weighted average help us analyse the model better as our training and test sets are uneven for different classes.



```
              precision    recall  f1-score   support

       anger       0.90      0.86      0.88       159
     disgust       0.84      0.79      0.81       155
        fear       0.86      0.84      0.85       164
       happy       0.72      0.82      0.77       144
         sad       0.83      0.86      0.84       149
   sarcastic       0.68      0.77      0.73       133
    surprise       0.90      0.75      0.82       146

    accuracy                           0.82      1050
   macro avg       0.82      0.81      0.81      1050
weighted avg       0.82      0.82      0.82      1050
```

fig 13) Classification Report

V. *CONCLUSION*

In this paper, we have proposed a model to recognize emotion from speech which is in Hindi. The model was able to analyze the tone and pitch of various audio clips of both, male and female voices, and classify them into emotions like angry, sad, happy, sarcastic, surprised and so on. The importance of the prosodic and spectral parameters for discriminating the emotions is shown by performing the emotion classification using prosodic and spectral features. The final accuracy obtained with MLP Classifier was 81.52%.

VI. *FUTURE WORK*

The proposed model can be further exploited for better efficiency if we have larger data set available to us in future.

Extension into real-world systems like customer service can be attempted.

REFERENCES

[1] S. G. Koolagudi, R. Reddy, J. Yadav and K. S. Rao, "IITKGP-SEHSC : Hindi Speech Corpus for Emotion Analysis," 2011 International Conference on Devices and Communications (ICDeCom), Mesra, 2011, pp. 1-5, doi: 10.1109/ICDECOM.2011.5738540

[2] Ashish B. Ingale, D. S. Chaudhari, "Speech Emotion Recognition" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2 Issue-1, March 2012

[3] A. V. Oppenheim, A. S. Willsky, and H. Nawab, Signals and Systems. Prentice Hall, 1996. [2] J. G. Proakis and D. G. Manolakis, Digital Signal Processsing. Prentice Hall, 1996.

[4] sklearn.neural_network.MLPClassifier, https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[5] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.

[6] *Verhulst, Pierre-François (1838). "Notice sur la loi que la population poursuit dans son accroissement". Correspondance Mathématique et Physique.* ***10****: 113–121.*

[7] How To Apply Machine Learning And Deep Learning Methods to Audio Analysis - https://hackernoon.com/how-to-apply-machine-learning-and-deep-learning-methods-to-audio-analyis-wt6p32qz

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[9] Sujay G. Kakodkar, Samarth Borkar, "Speech Emotion Recognition of Sanskrit Language using Machine Learning", in International Journal of Computer Applications (0975 – 8887) Volume 179 – No.51, June 2018

[10] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[11] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[12] K. Elissa, "Title of paper if known," unpublished.

[13] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[14] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[15] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.