Human Speech Emotion Recognition Using Machine Learning

Mrs. Sandhya Gundre
professor ,
D.Y Patil Institue of Engineering
management and research akurdi
Pune, India
sandhya.gundre@dypiemr.ac.in

Nabeel Khan
dept. of computer science
D.Y Patil Institue of Engineering
management and research akurdi
Pune, India
nabeel.aijazullah@gmail.com

Amisha Punjabi
dept. of computer science
D.Y Patil Institue of Engineering
management and research akurdi
Pune, India
amishapunjabi123@gmail.com

Raj singh
dept. of computer science
D.Y Patil Institue of Engineering
management and research akurdi
Pune, India
singhrj002@gmail.com

Rajat Arora
dept. of computer science
D.Y Patil Institue of Engineering
management and research akurdi
Pune, India
rajatarora19297@gmail.com

Abstract –The investigation of human discourse is a difficult examination territory as it concerns the discovery of user networks. Feelings assume an underlying part in human communication. The capacity to comprehend human feelings by breaking down voice is alluring in various uses of discourse acknowledgment in feelings can be found in various zones, for example, the association among PCs and people and call focuses. Already, feeling acknowledgment utilized straightforward classifiers on bag- of-words models. Nonetheless, the current work of feeling acknowledgment on Voice was done with the assistance of profound learning strategies on static voice information. The proposed strategy centers around expanding the general precision of feeling acknowledgment during calls utilizing man-made consciousness. The general point is to precisely perceive the different feelings that a specific speech communicates semantically.

.

Keywords- Human Emotion, feature extraction, machine learning, Speech conversion.

I. INTRODUCTION

Emotions can be defined as conscious affect attitudes, which constitute the display of a feeling. In recent years, a large number of studies have focused on emotion detection using opinion mining on speech. Due to some intrinsic characteristics of the voice produced during calls, such as the loudness, voice quality and casual expression, emotion recognition on them is a challenging task. Previous studies have focused mainly on lexical and deep learning methods. The performance of vocabulary-based methods largely depends on the quality of the emotional lexicon and the performance of deep learning methods depends largely on the characteristics. Therefore, we work with twoclassifiers that are the most famous, and have also been used before by the researchers from computational linguistics and natural language processing (NLP). Finally, Profile of Mood States is a psychological equipment that defines a four-dimensional mood state representation using text. The novel technique a Profile of Mood States generating Four-dimensional mood state representation using 65 adjectives with combination of emotions categories like, anger, happy, sad and normal.Previous work generally studied only one emotion recognition. Working with multiple classifications simultaneously not only enables performance comparisons between different emotion categorizations on the same type of data, but also allows us to develop a single model for predicting multiple classifications at the same time.

II. LITERATURE REVIEW

The fundamental thought of the paper [1] is to apply Deep Neural Network (DNN) and knearestneighbor (k-NN) in grouping of feeling from discourse particularly alarming perspective. The region of utilization of the framework is basically worried over the medical services units. The establishment of this exploration has its primary firm applications in palliative consideration. Under most exact result the alarm signals are made through cloud. Numerous crude information are gathered under unique accentuation procedures. The voice signals are changed over to wave structure, expression level element extraction feeling acknowledgment, and ready sign creation through cloud is the arrangement of steps to be followed.

In this paper [2], a Speech Emotion Recognition (SER) framework is proposed utilizing the component mix of Teager Energy Operator (TEO) and Linear Prediction Coefficient (LPC) highlights as T-LPC include extraction. The focused on discourse signals which were not precisely perceived in the past SER frameworks were perceived utilizing the proposed techniques. Gaussian Mixture Model (GMM) classifier is utilized to sort the feelings of EMO-DB information base in this examination. The Stressed Speech Emotion Recognition (SSER) proposed utilizing the T-LPC highlight extraction procedure procured better execution contrasted with the current Pitch, LPC, and LPC + Pitch include based acknowledgment frameworks. This proposed feeling acknowledgment framework can be utilized to inspire the understudies by finding their enthusiastic state giving better exactness contrasted with the current ones.

In this paper [3], another organization model (CNN-RF) in light of A neural convolution network joined with an arbitrary timberland is proposed. To begin with, the neuronal convolution network is utilized as an extractor of highlights to separate the qualities of vocal feelings from the standardized spectrogram, an arbitrary woods order calculation is utilized to arrange the attributes of vocal feelings. The aftereffect of test shows that CNN-RF model is better than the customary CNN model. Furthermore, Improved the Record Sound order box of Nao and applied the CNN-RF model to Nao robot.

The paper [4] builds up a perform various tasks DNN for learning assignments across different errands, not just utilizing tremendous measures of cross-task information, yet additionally profiting by a regularization impact that prompts more broad portrayals to help undertakings in new areas. A perform various tasks profound neural organization for portrayal learning, specifically zeroing in on semantic characterization (inquiry grouping) and semantic data recovery (positioning for web search) assignments. Show solid outcomes on inquiry order and web search. Focal points are: The MT-DNN unequivocally performs utilizing solid baselines across all web search and question characterization errands. Perform multiple tasks DNN model effectively joins assignments as unique as grouping and positioning. Inconveniences are: The inquiry arrangement consolidated either as grouping or positioning undertakings not thorough investigation work.

In article [5], show that feeling word hashtags are acceptable manual names of feelings in tweets. Proposes a strategy to produce a huge vocabulary of word feeling relationship from this feeling marked tweet corpus. This is the primary dictionary with genuine esteemed word feeling affiliation scores. Points of interest are: Using hashtagged tweets can gather a lot of named information for any feeling that is utilized as a hashtag by tweeters. The hashtag feeling vocabulary is performed essentially in a way that is better than those that utilized the physically made WordNet influence dictionary. Consequently recognizes character from text. Detriments are: This paper works just on given content not equivalent word of that text.

The paper [6] centers around contemplating two central NLP assignments, Discourse Parsing and Sentiment Analysis. The improvement of 3 free recursive neural nets: for the key sub-commitments of talk parsing, explicitly structure forecast and connection expectation; the 1/3 web for supposition expectation. Points of interest are: The idle Discourse highlights can help support the presentation of a neural estimation analyzer. Pre-preparing and the individual models are a significant degree quicker than the Multi-entrusting model. Hindrances are: Difficult expectations to multi-sentential content.

In this paper [7] they we present our discoveries on how the learning of portrayal in a huge plain vocal corpus can be utilized in a helpful manner for the acknowledgment of the feelings of language (BE). We show that the incorporation of the took in portrayals from an unattended programmed encoder into a CNN-based feeling classifier improves acknowledgment exactness.

In this paper [8], He proposed another model for the nonstop acknowledgment of feelings from language. This model, which has been prepared from one finish to the next, is made out of a convolutional neural organization (CNN), which extricates the attributes of the natural sign and stacks a momentary long haul 2-layer memory (LSTM), for Consider the logical data in the information.

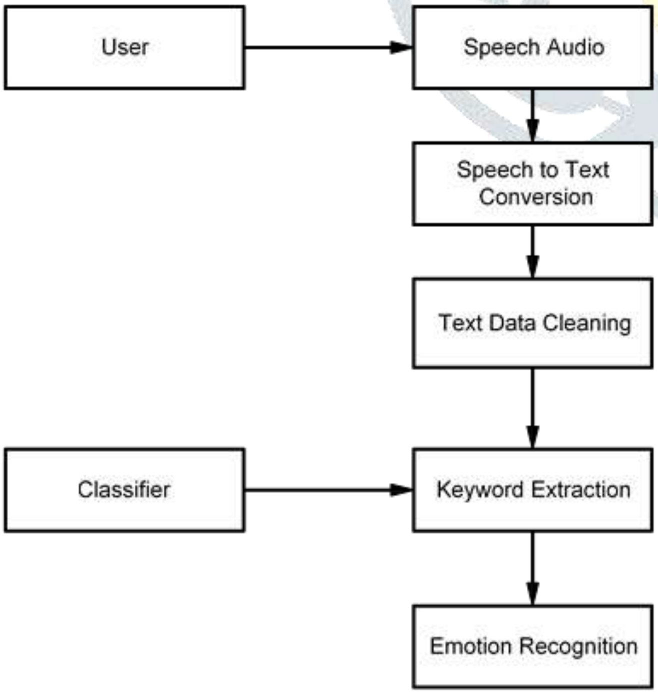
This paper [9] propose to incorporate the consideration system into profound intermittent neural organization models for discourse feeling acknowledgment. This depends on the instinct that it is useful to underscore the expressive piece of the discourse signal for feeling acknowledgment. By presenting consideration instrument, we cause the framework to figure out how to zero in on the more hearty or useful portions in the info signal. The proposed incorporation of consideration instrument on top of the gauge profound RNN model accomplishes 46.3% UA review rate.

In this paper [10] they utilize 13 MFCC (Mel Frequency Cepstral Coefficient) with 13 speed and 13 increasing speed segment as highlights and a CNN (Convolution Neural Network) and LSTM (Long Short Term Memory) based methodology for arrangement. We picked Berlin Emotional Speech dataset (EmoDB) for grouping reason. We have roughly 80% of precision on test information.

III. Proposed Approach

Text perceives from human discourse utilizing discourse transformation library through component extraction strategies. Human Mood States is a mental instrument for surveying the person's temperament state. It characterizes 65 descriptive words that are appraised by the subject on the five-point scale. Every descriptive word adds to one of the four classifications. The higher the score for the modifier, the more it adds to the general score for its classification, aside from loose and effective whose commitments to their individual classifications are negative. Disposition states consolidates these evaluations into a four-dimensional temperament state portrayal comprising of classifications: outrage, glad, tragic and typical. Contrasting with the first structure, we disposed of the modifier blue, since it just seldom relates to a feeling.

A. **System Architecture**



B. **Algorithm:**

1. Hidden Markov Model (HMM) algorithm for speech recognition:

A HMM is characterized by 3 matrices viz., A, B and PI.

A - Transition Probability matrix $(N \times N)$

B - Observation symbol Probability Distribution matrix $(N \times M)$

PI - Initial State Distribution matrix $(N \times 1)$

Where, N =Number of states in the HMM

M = Number of Observation symbols

After can apply HMM for speech recognition by using following steps:

1. Recursive procedures like forward and Backward Procedures exist which can compute P (O|L), probability of observation sequence.

Forward Procedure:

Initialization:

$$\alpha_1(i) = \pi_i b_i o_1, \qquad 1 \le i \le N$$

Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}),$$

$$1 \le t \le T-1, 1 \le j \le N$$

Termination

$$P(O|\lambda) \sum_{i=1}^{N} \alpha_T(i)$$

Backward Procedure:

Initialization:

$$\beta_T(i) = 1, \qquad 1 \le i \le N$$

Induction

$$\beta_T(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1}) \beta_{t+1}(j),$$

$$T-1 \le t \le 1, 1 \le i \le N$$

Termination

$$P(O|\lambda) \sum_{i=1}^{N} \alpha_T(i)$$

2. The state occupation probability $t(sj)$ is the probability of occupying state sj at time t given the sequence of observations

$$O_1, O_2, \dots, O_N.$$

3. Baum-welch algorithm for parameter re-estimation.

2. MMLDA Algorithm for summarization

Steps:

1. For the topic T, draw $\varphi^{TG}\sim Dir(\lambda^{TG})$and $\varphi^{VG}\sim Dir(\lambda^{VG})$ denote the general textual distribution and visual distribution, respectively. $Dir(\cdot)$is the Dirichlet distribution. Then draw $\phi^{Z} \sim Dir(\beta^{Z})$, which indicates the distribution of subtopics over the microblog collection corresponding to T.
2. For each subtopic, draw $\varphi_K^{TS} \sim Dir(\lambda^{TS})$and $\varphi_K^{VS} \sim Dir(\lambda^{VS})$, $k \in \{1, 2, \ldots, K\}$, correspond to the specific textual distribution and visual distribution.
3. For each microblog M_i, draw $Z_i \sim Multi(\phi^{Z})$, corresponds to the subtopic assignment for Mi. Multi(·) denotes the Multinomial distribution. Then draw $\phi_i^{R} \sim Dir(\beta^{R})$ indicates the general-specific textual word distribution of M_i. Similarly, draw $\phi_i^{Q}\sim Dir(\beta^{Q})$ indicates that for visual words.
4. For each textual word position of M_i, draw a variable $R_{ij} \sim Multi(\phi_i^{R})$:
    - If R_{ij} indicates General, then draw a word $W_{ij} \sim\ Multi(\varphi^{TG})$.
    - If R_{ij} indicates Specific, draw a word W_{ij} from the Z_i-th specific distribution $W_{ij} \sim Multi(\varphi_{Z_i}^{TS})$
5. The generation of visual words is similarly done as in step 4.

IV. RESULTS AND DISCUSSION

The experimental result evaluation, we have notation as follows:

TP: True positive (correctly predicted number of instance)

FP: False positive (incorrectly predicted number of instance),

TN: True negative (correctly predicted the number of instances as not required)

FN false negative (incorrectly predicted the number of instances as not required),

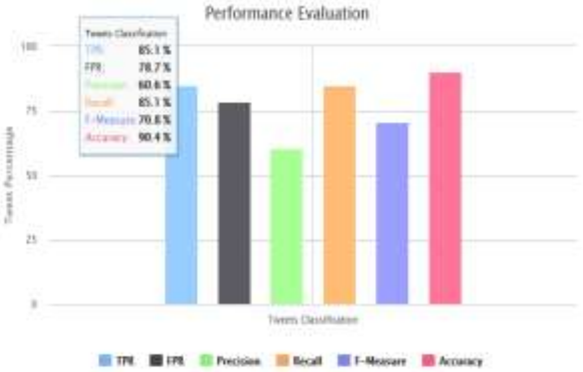
On the basis of this parameter, we can calculate four measurements

Accuracy = TP+TN÷TP+FP+TN+FN

Precision = TP ÷TP+FP

Recall= TP÷TP+FN

F1-Measure = 2×Precision×Recall ÷Precision+ Recall.



Parameters	Percentage
TPR	85.1
FPR	78.7
Precision	60.6
Recall	85.1
F-Measure	78.8
Accuracy	94.4

Conclusion

The method of recognizing the temperament or state of an individual through voice is an arising thought where the helpfulness of this cycle is inescapable, and will impart its uses to numerous areas from clinical to data advancements. This venture actualizes an Emotion Recognition on Speech utilizing novel techniquea Profile of Mood States (POMS)using multinomial innocent Bayes speaks to four-dimensional temperament state portrayal utilizing 65 modifiers with blend of feelings classifications like upbeat, miserable, outrage and typical.

REFERENCES

[1] K.Tarunika , R.B Pradeeba , P.Aruna” Applying Machine Learning Techniques for Speech Emotion Recognition” ICCCNT 2018.

[2] Surekha Reddy B, T. Kishore Kumar” Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features” 2018 IEEE 18th International Conference on Advanced Learning Technologies

[3] Li Zheng, Qiao Li2, Hua Ban , Shuhua Liu1” Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest”IEEE 2018

[4] O. Irsoy and C. Cardie, “Opinion Mining with Deep Recurrent Neural Networks,” in Proc. of the Conf. on Empirical Methods in Natural Language Processing. ACL, 2014, pp. 720–728.

[5] S. M. Mohammad and S. Kiritchenko, “Using Hashtags to Capture Fine Emotion Categories from Tweets,” Computational Intelligence, vol. 31, no. 2, pp. 301–326, 2015.

[6] B. Nejat, G. Carenini, and R. Ng, “Exploring Joint Neural Model for Sentence Level Discourse Parsing and Sentiment Analysis,” Proc. of the SIGDIAL 2017 Conf., no. August, pp. 289–298, 2017.

[7] Michael Neumann, Ngoc Thang Vu,” IMPROVING SPEECH EMOTION RECOGNITION WITH UNSUPERV”IEEE 2019.

[8] Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller “END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS”IEEE 2018

[9] Po-Wei Hsiao and Chia-Ping Chen” EFFECTIVE ATTENTION MECHANISM IN DYNAMIC MODELS FOR SPEECH EMOTION RECOGNITION”IEEE 2018

[10] SaikatBasu, Jaybrata Chakraborty, Md. Aftabuddin” Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural Network Architecture” International Conference on Communication and Electronics Systems (ICCES 2017)