Kannada Speech Emotion Recognition Using Ensembling Techniques

SMRITHI BALIGA[1], SAPNA H M[2], SHREYAS N[3], YOGESH GOWDA V[4], DR CHANDRASHEKAR M PATIL[5], PROF. AUDRE ARLENE[6]

[1, 2, 3, 4, 5] Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, India

[6] Assistant Professor, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, India

Abstract- This study explores the development of a speech emotion recognition system for the Kannada language, using a dataset of audio recordings labeled with six emotion categories: happiness, sadness, anger, fear, and neutral. We used a combination of acoustic features and machine learning algorithms, including Mel-frequency cepstral coefficients (MFCCs), to classify emotions in the audio recordings. Our results show that the proposed system achieves an average accuracy of 75% on the Kannada emotion dataset, outperforming existing baseline models. These findings suggest that Kannada speech emotion recognition can be achieved with high accuracy using a combination of acoustic features and machine learning algorithms like RNN, CNN and DBN, paving the way for further research in this area.

Indexed Terms- Speech Emotion Recognition, Mel-Frequency Cepstral Coefficients, Recurrent Neural Network, Deep Belief Network

I. INTRODUCTION

Feeling distinguishing proof from discourse has created from a particular field to a significant component in Human-PC Communication (HCI). These frameworks expect to work with the regular connection with machines by direct voice collaboration as opposed to involving customary gadgets as contribution to comprehend verbal substance and make it simple for human audience members to respond. A few applications incorporate discourse frameworks for communicated in dialects, for example, call focus discussions, on board vehicle driving framework and use of feeling designs from the discourse in clinical applications. Regardless, there are numerous issues in HCI frameworks that actually should be appropriately tended to, especially as these frameworks move from lab testing to true application. Consequently, it is expected to successfully take care of such issues and accomplish better feeling acknowledgment by machines.

II. PROPOSED METHODOLOGY

A unique answer to Kannada speech emotion recognition is the "Kannada Speech Emotion Recognition Using Ensembling Techniques". Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. First and foremost, we use MFCC for include extraction. The Mel-recurrence Cepstral Coefficients highlight extraction procedure essentially incorporates windowing the sign, applying the DFT, taking the log of the size, and afterward distorting the frequencies on a Mel scale, trailed by applying the reverse DCT. Next ensembling procedures are utilized for feeling acknowledgment. Gathering learning strategies can be utilized on brain organizations to work on the presentation of discourse feeling acknowledgment errands. Here use RNN and DBN are utilized for ensembling. A repetitive brain organization (RNN) is a sort of brain network ordinarily utilized in discourse feeling acknowledgment. RNNs are intended to perceive the successive attributes in information and use examples to foresee the following likely situation. Profound Conviction Organization (DBN) can be utilized to address solo learning errands to decrease the dimensionality of discourse and close to home highlights. This is the functioning clarification of our proposed framework.

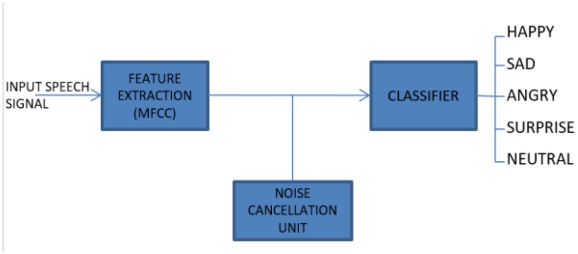

Figure 1: Work Flow

III. LITERATURE REVIEW

M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju [1] Speech emotion recognition is a technique proposed, which extracts a speaker's emotional state from their speech. The voice signal's acoustic feature is called Feature. A little amount of information from the speech signal is extracted through the feature extraction method so that it may later be utilised to identify each speaker. MFCC, one of many component extraction procedures, is a generally used strategy. The information taken from the speaker discourse signal is utilized to recognize the speaker's states of mind. A speaker's inclination can be recognized from their discourse utilizing the MFCC approach. The created framework has been tried for happiness, sadness, rage, emotions and the productivity was viewed as around 80%.

Anagha Sonawane, M U Inamdar, Kishor B Bhangale [2] have suggested that one of the difficulties in speech processing and Human Machine Connection (HMI) for the point of settling different functional targets for this present reality applications, is the improvement of a framework for feeling acknowledgment utilizing human discourse. Talking is a characteristic method for communicating sentiments that offers profundity. In this paper, the Different Help Vector Machine (SVM) is utilized as a classifier and the MFCC is utilized to extricate highlights. The library of happy, angry, sad, disgusted, surprised, and neutral emotion sounds inclination sounds has likewise been the subject of serious trial and error. Various SVM's presentation examination showed that non-direct bit SVM beat straight SVM regarding precision.

Demircan Semiye & Kahramanli Humar [3] proposed a voice and emotion recognition system. In this paper, pre-handling necessaries for feeling acknowledgment from discourse information, have been performed. The explores in the space have been demonstrated that significant outcomes have acquired utilizing prosodic elements of discourse. To perceive feeling a few prosodic elements have been extricated first (Measurable information separated from F0), second Mel Recurrence Cepstral Coefficients (MFCC) and thirdly LPC (straight forecast coefficients). Separated highlights have characterized with ANN (Fake Brain Organization), SVM (Backing Vector Machines), kNN (k-Closest neighbor Calculation).

S. Mirsamadi, E. Barsoum and C. Zhang [4] proposed a voice and feeling acknowledgment framework. Alongside voice acknowledgment, conversational perspectives including song, feeling, tone, and accentuation are being explored. Research has shown the way that prosodic parts of discourse can be utilized to accomplish critical advantages. In this review, the pre-handling expected for voice information feeling acknowledgment was completed. From the voice signal, attributes have been recovered. MFCCs, which were taken from the signs and sorted utilizing the k-NN technique, were utilized to recognize feeling

Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, Zhe Wang [5] fostered a strategy Bi-directional Long-Transient Memory with Directional Self-Consideration (BLSTM-DSA), a procedure that was intended to mine the significance of signs in sounds and grow the variety of data, is recommended in this exploration. From learnt nearby elements, the Long Momentary Memory (LSTM) can gather long haul conditions. Furthermore, Bi-directional Long-Momentary Memory (BLSTM) can reinforce the design through a heading system since it can all the more precisely recognize covered feelings in sentences. Also, the absence of data can be tended to by utilizing autocorrelation of discourse outlines, which considers the presentation of the Self-Consideration instrument into SER.

Rather not entirely set in stone by adding the consequences of the forward and in reverse LSTM, the consideration weight for each edge is resolved utilizing their unmistakable results. Thus, the program can precisely pick discourse outlines in a fleeting organization that incorporate close to home data via consequently commenting on the loads of discourse outlines. The BLSTM-DSA performs acceptably on the errand of discourse feeling acknowledgment when tried against the Intelligent Close to home Dyadic Movement Catch (IEMOCAP) data set and Berlin information base of profound discourse (Emotional DB). The most extreme acknowledgment precision is accomplished by BLSTM-DSA, especially in the acknowledgment of bliss and outrage.

Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder [6] a methodology that looks at discourse feeling acknowledgment frameworks has been introduced. Present are a hypothetical definition, a classification of full of feeling state, and the different ways that feelings can be communicated. A SER framework based on different classifiers and elements extraction procedures is made to do this review. The voice signals are handled to extricate MFCC and balance otherworldly (MS) attributes, which are then used to prepare different classifiers. To track down the most appropriate element subset, include determination (FS) was utilized. For the feeling characterization challenge, various AI ideal models were utilized. Seven feelings are at first ordered utilizing a RNN classifier. Their outcomes are a short time later diverged from those of multivariate straight relapse (MLR) and support vector machine (SVM) strategies, which are regularly utilized in the field of spoken sound sign feeling acknowledgment. The exploratory information assortment comprises of the Berlin and Spanish data sets. This review shows that when speaker standardization (SN) and element determination are applied to the highlights, all classifiers for the Berlin information base arrive at an exactness of 83%. RNN classifiers without SN and with FS accomplish the best exactness (94%) for Spanish information bases.

Bu Chen, Qian Yin, Ping Guo [7] here a DBN structure was introduced as a profound learning approach application to the extraction of feelings present in Chinese discourse. Instead of traditional shallow learning methods, a DBN classifier is utilized to perceive feelings from mandarin discourse utilizing eight fitting factors, including pitch and the MFCC. Here, it has been shown that the acknowledgment rate is higher than that of the SVM classifier and the customary back spread (BP) approach.

Haiqing Zheng, Yaru Yang [8] developed a model for voice feeling acknowledgment in view of a superior DBN. The Corrected Direct Unit is utilized instead of the ordinary DBN actuation capability in this strategy, and the reproduction blunder is utilized to appraise the profundity of the DBN organization. As the center components of the profound discourse signal, the brief time frame energy, brief time frame zero intersection rate, essential recurrence, formants, and 24 layered MFCC boundaries are recovered. Programmed acknowledgment of the six feelings — outrage, dread, euphoria, serenely, distress, and shock — is achieved involving these key attributes as contribution to the DBN. The changed DBN model accomplishes a superior acknowledgment result when contrasted with the traditional DBN model and the BP model, and the acknowledgment rate is expected to reach 84.94%.

Peng Shi [9] describes well-known speech emotion databases and proposes a method for recognising speech emotions using a continuous model and a discrete model. To give a more precise description of speech emotion, this essay looks at a variety of characteristics. The main contributions of this research were the selection of the database, the extraction of emotion features, and the selection of a classification method. The overall and average recognition rate indicators are then used to evaluate the outcomes. In this research, emotion-related characteristics are extracted via contrastive divergence. The accuracy of the test emotion sample performs better after feature extraction by DBN, which is given as being 5% higher than standard classification methods, such as support vector machine (SVM) and artificial neural networks.

V B Kobayashi, V B Calag [10] It is recommended in this research that ensemble learning methods be utilised to create classifiers such random forest and kernel factory that are employed in the development of a speech emotion recognition system. Pre-processing voice samples, feature extraction, classifier building, and ultimately emotion prediction are all steps the system takes. Fundamental frequency, energy, MFCCs, and linear predictive cepstrum coefficients characteristics were extracted from each segment. They also trained two ensemble classifiers, random forest and kernel factory, to test the approach using various speech databases. The results showed that ensemble classifiers only outperformed single models in classification by a maximum of 20%. Ensemble classifiers are suitable for this field since they are effective at recognising emotions.

N. T. Ira and M. O. Rahman [11] The goal of this study is to recognise different emotions in audio speech. Here, emotions are categorised into eight separate groups: neutral, fear, pleasure, anger, sadness, disgust, quiet, and astonishment. To extract features, Mel Frequency Cepstral Coefficients have been used (MFCCs). Multilayer perceptron (MLP), Random Forest (RF), AdaBoost, support vector machine (SVM), Gradient Boosting (GB), and Hist Gradient Boosting are the six main supervised classifier types that have been used for classification (HGB). This study proposes a novel method for identifying speech emotions. The accuracy rates for MLP, AdaBoost, SVM, Random Forest, Gradient Boosting, and Hist Gradient Boosting are 53%, 32%, 54%, 58%, 56%, and 59%, respectively. An accuracy rate of 70% was achieved after additional testing utilising the ensemble technique, which combines the RF, GB, and HGB. On the basis of a comparison analysis of six classifiers and other existing approaches, it is shown that the Ensemble Method is one of the best alternatives for emotion recognition from speech.

D. Valles and R. Matin [12] In this study, a speech emotion recognition system was developed with the intention of helping kids with ASD comprehend their communication partner's feelings better. This system was created using machine learning and deep learning techniques. The final prediction on the recorded input utterances was made using a variety of machine learning techniques using ensemble learning. This class of models consists of Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and Recurrent Neural Networks (RNN). This work covers the audio processing of the samples, approaches for including background noise, and feature extraction coefficients considered for model construction and training. This study presents the performance assessment of each individual model for each dataset, the addition of background noise, and the combination of using all of the samples in all three datasets. This ensemble learning outperforms the MLP model's 65.7% performance accuracy in classifying emotions with a performance accuracy of 66.5%.

Paper name	Keypoints	Paper name	Keypoints
Speech based human emotion recognition using MFCC	MFCC technique is used to recognize emotion of a speaker from their voice. Accuracy= 80%	An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Network	Replaced the traditional DBN activation function with a Rectified Linear Unit(Relu). Speech emotion recognition rate is about 84.94 %.
Sound based human emotion recognition using MFCC & multiple SVM	MFCC = extraction of features and Multiple SVM as a classifier.	Speech Emotion Recognition Based on Deep Belief Network	Used contrastive divergence algorithm on emotion feature extraction.
Feature Extraction from Speech Data for Emotion Recognition	Performed pre-processing necessary for emotion recognition from speech data. Features	Detection of Affective States from Speech Signals using Ensembles of Classifiers	System is highly feasible since they employ an automatic segmentation process and the use of ensemble methods

	are extracted using MFCC		makes the training of the classifier fast and efficient.
Automatic SER using recurrent neural networks with local attention	Deep learning to automatically discover emotionally relevant features from speech.	An Efficient Speech Emotion Recognition Using Ensemble Method of Supervised Classifiers	Supervised classifiers - multilayer perceptron (MLP), RF, SVM have been used for classification Accuracy : 70%
Automatic Speech Emotion Recognition Using Machine Learning	RNN is used first to classify seven emotions. Berlin database - 83% Spanish database-94%	An Audio Processing Approach using Ensemble Learning for Speech-Emotion Recognition for Children with ASD	Ensemble of models includes a SVM, MLP, and RNN. All three models were trained on the RAVDESS, the TESS.
A Study of Deep Belief Network Based Chinese Speech Emotion Recognition on MFCC & Multiple SVM	A DBN classifier is used. Speech emotion recognition accuracy rate is about 87.5 %.	Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Systems with Applications	Ability to process short-term spectral features but yet respond to long-term temporal events.

CONCLUSION

This paper has given a nitty gritty survey of the profound learning strategies for MFCC, DBN, RNN and furthermore about ensembling procedures. profound learning strategies, for example, rnn, dbn have been the subject of much exploration as of late. these profound learning techniques and their layer-wise structures are briefly expounded in view of the classification of different normal feeling like bliss, happiness, bitterness, unbiased, shock, weariness, revulsion, dread, and outrage. these techniques offer simple model preparation and furthermore has greatest productivity rates. this writing work shapes a base to assess the presentation and impediments of current profound learning procedures. further we can presume that productive use of MFCC for discourse signal extraction, likewise classifiers, for example, RNN, DBN procedures for discourse signal groupings including a few other proficient classifiers can be utilized for accomplishing higher precision rates.

REFERENCES

[1] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET),2017,pp.2257-2260,doi: 10.1109/WiSPNET.2017.8300161.

[2] Sonawane, Anagha et al. “Sound based human emotion recognition using MFCC & multiple SVM.” 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC) (2017): 1-4.

[3] Demircan, Semiye & Kahramanli, Humar. (2014). Feature Extraction from Speech Data for Emotion Recognition. Journal of Advances in Computer Networks. 2. 28-30. 10.7763/JACN.2014.V2.76.

[4] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing

(ICASSP), 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.

[5] Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, Zhe Wang,"Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Systems with Applications", Volume 173, 2021, 114683, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.114683.

[6] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning", in Social Media and Machine Learning. London, United Kingdom: IntechOpen, 2019 [Online]. Available: https://www.intechopen.com/chapters/65993 doi: 10.5772/intechopen.84856

[7] P. Shi, "Speech emotion recognition based on deep belief network," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, pp. 1-5, doi: 10.1109/ICNSC.2018.8361376.

[8] B. Chen, Q. Yin and P. Guo, "A Study of Deep Belief Network Based Chinese Speech Emotion Recognition," 2014 Tenth International Conference on Computational Intelligence and Security, 2014, pp. 180-184, doi: 10.1109/CIS.2014.148.

[9] H. Zheng and Y. Yang, "An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Network," 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2019, pp. 493-497, doi: 10.1109/ICPICS47731.2019.8942482.

[10] V. B. Kobayashi and V. B. Calag, "Detection of affective states from speech signals using ensembles of classifiers," IET Intelligent Signal Processing Conference 2013 (ISP 2013), 2013, pp. 1-9, doi: 10.1049/cp.2013.2067.

[11] N. T. Ira and M. O. Rahman, "An Efficient Speech Emotion Recognition Using Ensemble Method of Supervised Classifiers," 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), 2020, pp. 1-5, doi: 10.1109/ETCCE51779.2020.9350913.

[12] D. Valles and R. Matin, "An Audio Processing Approach using Ensemble Learning for Speech-Emotion Recognition for Children with ASD," 2021 IEEE World AI IoT Congress (AI IoT), 2021, pp. 0055-0061, doi: 10.1109/AIIoT52608.2021.9454174.