Use of Artificial Intelligence in Emotion Recognition by Ensemble based multilevel classification

Sandeep Rathor[1]*, Megha Kansal[2], Mansi Verma[3], Madhav Garg[4] and Rishabh Tiwari[5]
[1,2,3,4,5]Department of Computer Engineering & Applications
GLA University, Mathura
[1]*sandeep.rathor@gla.ac.in*

Abstract: Emotion is a part of communication. By the communication we can recognize the emotions whether the communicator is happy, sad, surprise, etc. The objective of this research is to decrease the gap between humans and machines taking emotions into account. In this regard, an artificial intelligence with machine learning techniques is used. In this paper, 8 distinct emotions are used as happy, sad, surprised, angry, calm, neutral, disgust and fearful. For feature extraction, MFCC, chroma and mel is used. The proposed model is executed on the standard LibriSpeech dataset. For feature reduction principal component analysis is used before train the model. The classification is also done by using the basic machine learning classifiers like KNN, SVM and MLP and found the individual accuracy of the classifiers 29.83%, 68.69% and 61.50% respectively. To enhance the accuracy more, we propose an ensemble based classification model using the MLP Classifier with 300 hidden layers, SVM (rbf) and KNN (weight uniform, distance), AdaBoost and Gradient Boost classifier. The obtained accuracy of the proposed model is higher than any of the individual classifier i.e. 90%.

Keywords: Use of Artificial Intelligence, Ensemble based artificial intelligence, Machine learning techniques, Emotion Recognition.

1. Introduction

Speech is the act of communication and expression of thoughts and feelings by spoken words. It is the most natural or easiest way to express ourselves [1]. In today's world, emojis have become very common in text messages because text messages could be misunderstood sometimes, and we would like to express or pass our emotions along with text as same we do in speech [1]. So, emotions help us to understand each other in a better way.

Emotions are basically a class of feelings. Research has revealed the significant role that emotions play in shaping human-social interaction [2]. The emotional detection is easy for humans, as it can be easily understood by humans but it is very difficult for machines to detect emotions [3]. Therefore, speech emotion detection refers to a set of objectively measurable parameters in voice that reflect the affective state a person is currently expressing so that the interaction between human and machine will involve an emotional feedback framework [3]. Machines could help people to take right decisions by detecting emotions. The emotion expression or visualization depends on the vocal features, facial expressions, postures, body movements, environment and culture of the speaker. The speaking style of the speaker also gets change as the environment and the culture gets change [4]. There are different types of emotions used by the speaker, depending on the way we interact or have an influence on any situation- happiness, anger, fear, surprise, sadness, disgust, etc.

We already use speech recognition in our everyday life such as Google speech to text searching assistant. Similarly, Speech Emotion Recognition (SER) could be used to detect our emotions as well [1]. This area has received increasing number of research interest throughout the current years [2]. SER is a technology that extract emotional features from the speech signal [5]. It is generally composed of three main parts- (i) speech signal acquisition, (ii) feature extraction, and (iii) emotion recognition [5]. A classifier is an algorithm that maps the input data to a specific category. There are different types of classifiers that are available such as: Logistic Regression, Naïve Bayes, K-nearest neighbors (KNN), Artificial Neural Network (ANN), Recurrent Neural Networks (RNN) [6], and many others.

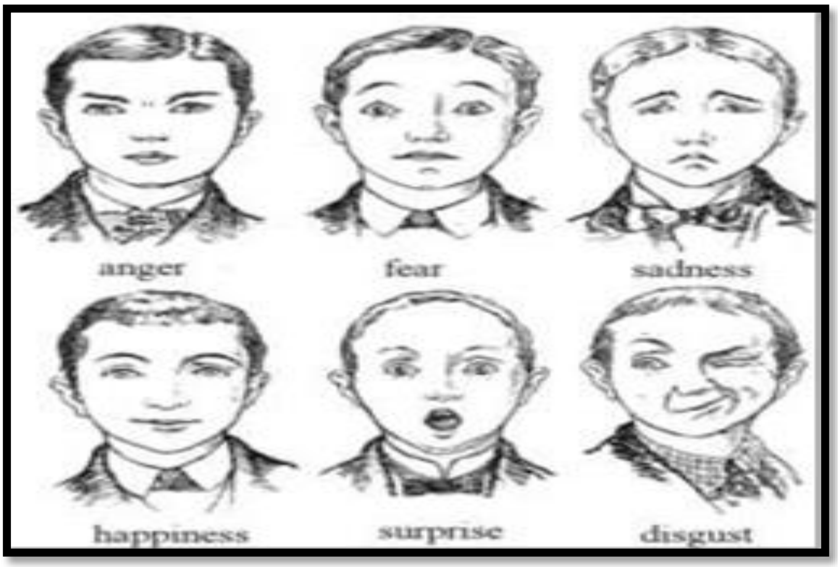

Figure 1: Different emotions of a person [1].

Rest of the paper organize as; in section 2 related work is discussed in the same context, section 3 designate the proposed model, section 4 discuss the results and discussion, and lastly section 5 depicted the conclusion and future aspects of the proposed model.

2. Related Work

In past years, different techniques have been proposed for speech emotion recognition. Various classifiers have been used for the classification of speech features such as Gaussian Matrix Model (GMM), Support Vector Machine (SVM), k-nearest neighbors (KNN), Artificial Neural Network (ANN), etc. Renjith et al., have worked on Telugu and Tamil languages to detect emotions happiness, sadness and anger using speech recordings [7]. In their work they have pre-processed to separate disturbances from speech waveforms and raw speech signals. They have extracted Hurst and Linear Predictive Cepstral Coefficients (LPCC) features and then classification is done on the basis of statistical parameters obtained from these features. Both KNN and ANN is used to identify the reactive emotions and then accuracy, precision and recall parameter is used to compare their performance for both features individually and in combination. The paper shown that Hurst gives better results than LPCC. However, the classes of emotions used in the research was very limited. Atreyee Khan, et al., have used Naive Bayes classifier along with both spectral and prosodic features for emotion detection [8]. As spectral features a Mel-Frequency Cepstral Coefficients (MFCC) has been used and pitch is used as prosodic feature. Naïve Bayes Classifier is used to perform classification and they have considered seven emotional classes to develop both gender independent and dependent system. Berlin Emo-db popular speech database speech samples are used to test accuracy of the system after performing classification. Energies spectrum of Discrete cosine transform (DCT) is used to calculate MFCC in which they have considered only 1-14 coefficients of DCT and rest is discarded. The execution of the research was well but the time complexity for training is very high. Rajasekhar et al., have distinguished the sample using SVM and speaker utterance is detected using MFCC [9]. In the end, SVM classifier differentiates between fear, anger, sadness, happiness and updates the database accordingly. The proposed approach is evaluated for combination of features in terms of accuracy that shows a good result for speaker independent cases as compared to individual features. The author done a great job however the dataset they have used, have poor quality audio samples with limited in numbers. Zheng et al., have proposed a random forest and CNN based new network model [10]. From normalized spectrogram a speech emotion features are extracted using CNN and then speech emotion features are classify using RF classification algorithm. From results it has been predicted that as compared to traditional CNN model use of CNN-RF model gives improved results. A good work has been done by him but has very few classes of emotions. In 2019, the author proposed an emotion recognition model [11]. In this model the author used both verbal and non-verbal sounds. Non-verbal sounds are laughter, cries etc. Firstly, the author developed the SVM based verbal/non-verbal detector. Then a prosodic phrase auto-tagger was used to extract the verbal/non-verbal part. For each part Convolutional neural network was used to extract the emotions and the sound features and then it is combined to form a CNN based generic feature vector. The accuracy obtained on the recognition of seven emotions state in NNIME is 52% which is very less.

A research on the combination of audio and text was proposed in 2019 [12]. In this model author processed text and audio both. In this proposed work, authors formed a binary model by combining LSTM and CNN. In the proposed model, LSTM network is utilized to capture the emotion feature and CNN is used to classify the fusion features. However, the accuracy of the proposed model is not up to the mark. In another model of emotion recognition, the author used two different datasets such as LDC (Linguistic Data Consortium) and UGC dataset to categories the emotions like sad, anger, happy, fear. In this paper, MFCC and LPCC are used to extract the best feature subset [13]. The accuracy obtained is 72.85% however, emotion classes are limited.

The objective of our paper is to increase the classes of emotions along with a good accuracy.

3. Proposed Model

We as a human can identify the emotion of the person through the communication. So, for a machine we need to define the parameters through which it can detect the correct emotion of a person. Therefore, we are going to propose a model which will detect different emotions from the communication with the higher accuracy.

This paper contains 12 males and 12 female professional actors. The dataset contains 24 actors speaking in 8 different emotions. Feature extraction is done through MFCC (Mel-frequency cepstral coefficients), MEL (Mel Spectrogram Frequency) and chroma. For classification purpose we use MLP, KNN, SVM AdaBoost and Gradient Boost classifiers. The complete process is shown in the figure 2.

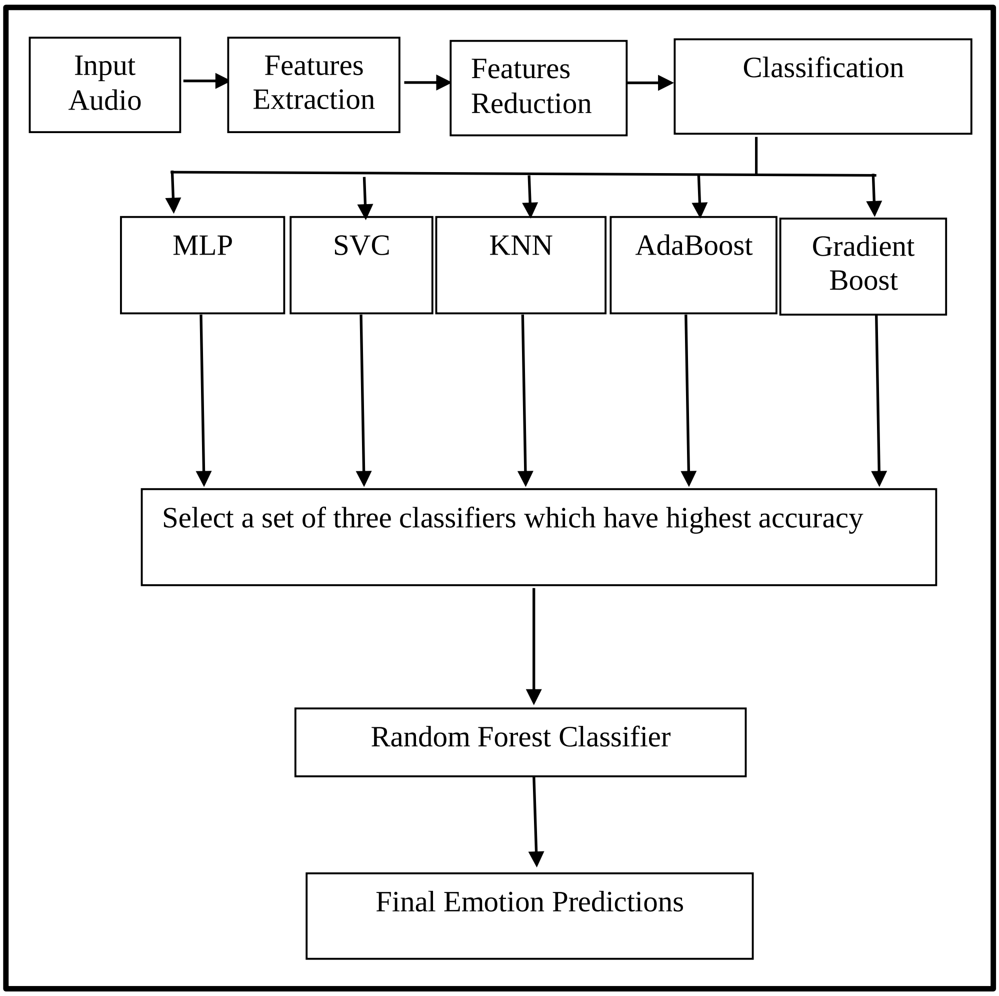

Figure 2: Proposed ensemble based multilevel frame work for emotion classification

There are various modules used in the proposed model.

- *Audio Feature Extraction*

We used librosa library to extract the speech features. A speech signal contains a very large number of parameters that reflect an emotional characteristic. In our proposed work, we have selected Mel-frequency cepstrum coefficients (MFCC), Mel Spectrogram Frequency (MEL) and chroma to extract the emotional features. The extraction of features aims to minimize data by transforming the input signal into a compact set of parameters while retaining the spectral and temporal features of the information of the speech signal.

- *Features Reductions and Features Selection*

To reduce the features, we applied Principal Component Analysis (PCA). It ensures there must be minimal loss of information and it decreases the computation time.There are so many features we extracted but there must be some features that would not affect the target so, we need to reduce the number of features used to characterize a dataset to improve the performance of the algorithm. Affective feature selection is the concept that a speech with a certain extracted features can express a corresponding emotion.

- *Classification*

To classify the emotion, initially five classifiers with different configurations are used. After that select the set of three classifiers which have the highest accuracy and it is passed to another classifier i.e. random forest to train the system. After the training, actual prediction is given by the random forest classifier [15]. By using an ensemble based approach our model works efficient and give the correct prediction with the accuracy of 90%.

4. Result and Discussion

The performance of the proposed ensemble based classification is discussed in this section. For the identification of emotion, the different samples are tested. When the samples are tested, the time and amplitude waveform generated as it is shown in the figure 3. The features which are the best suited, provided to the classifier after the extraction of features. A classifier is an algorithm that maps the input data to a specific category. Different types of classifier have been proposed for the detection of speech emotion [13]

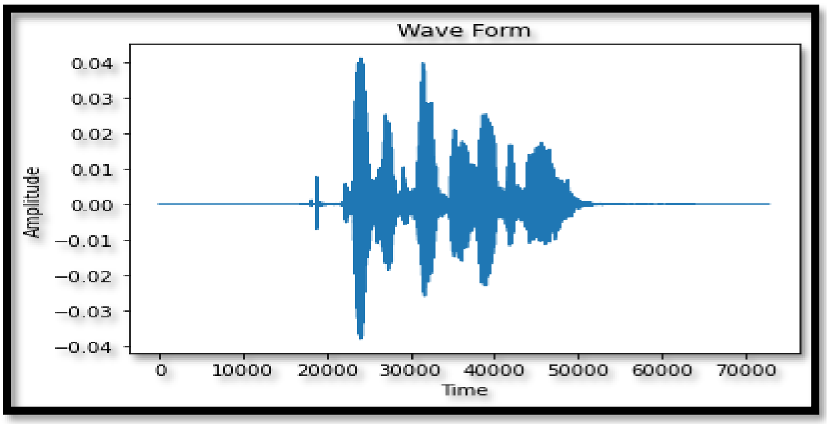

Figure 3: Waveform of time and amplitude

4.1 Used Dataset:

LibriSpeech dataset is used in this paper for training and testing purpose. The data in this dataset has been taken from audiobooks of the LibriVox project [14]. Total 2484 classes are there in LibriSpeech. The speech contains various emotions like happy, sad, fearful, angry, disgust, surprised, calm, etc.

4.2 Quantitative Analysis:

To evaluate the performance of the proposed model a quantitative analysis is shown in the table 1.

Table1: Testing and training accuracy of individual classifier

Classifier Configuration	Training Accuracy (%)	Testing Accuracy (%)
MLP Classifier	98.5%	61.50%
SVM Classifiers		
Kernel=linear	98.49%	58.93%
Kernel=rbf	97.47%	68.69%
Kernel=poly, degree=4	97.87%	39.29%
KNN (Nearest neighbor k=5)	87.42%	29.83%
AdaBoost	93.42%	69.98%
Gradient Boost	93.25%	68.87%

Table 1 represents different classifiers with configurations of hyper parameters to cover up all the instances of the data set. These classifiers with different configurations forms a set of classifiers that is used to train another classifier at higher level.

Table2: Accuracy of set of three classifiers

Classifiers Set	Highest Accuracy
SVC (linear), AdaBoost, MLP	88%
SVC (rbf), KNN, Gradient Boost	86%
MLP, SVC (polynomial), Gradient Boosting	84%

Table 2 represents the set highest accuracy of three classifiers. And it is found that SVC (linear), AdaBoost and MLP have the highest accuracy. Now, the prediction of these classifiers are passed to the next level to train the another classifier.

	Angry	happy	neutral	sad	disgust	fearful	calm	surprised
Angry	91	2	1	1	1	3	0	2
happy	1	89	2	0	1	2	1	2
neutral	1	3	91	1	0	3	0	1
sad	2	1	0	92	3	1	1	1
disgust	3	1	2	1	89	1	2	1
fearful	1	2	1	1	1	87	3	2
calm	1	1	1	3	3	2	92	2
surprised	0	1	2	1	2	1	1	89

Figure 3: Confusion matrix of emotion recognition on test data

Confusion matrix is shown in the figure 3. It shows the correct and incorrect classifications. The diagonal values represent the true positives or the correct classification while the others represent the misclassification. The accuracy can be obtain by the sum of diagonal, divided by the total instances and multiply by 100. i.e. 720/800*100=90%.

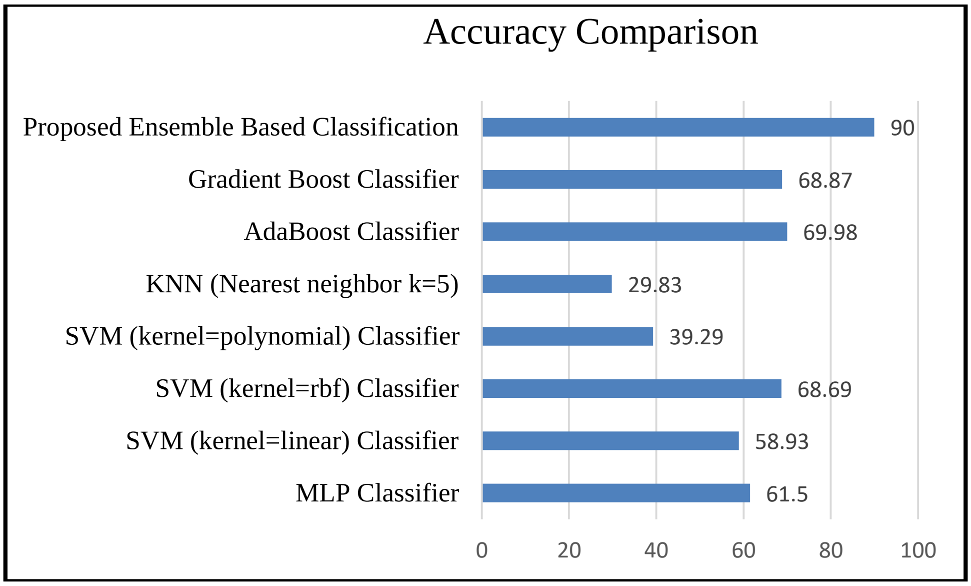

Figure 4: Accuracy comparison of proposed ensemble classification with others

The comparison of accuracy of the proposed ensemble based classification model with other existing models are shown in the figure 4. It is evident that our proposed model works better than any of the other existing models.

5. Conclusion and Future Scope

In the above study we performed the Speech Emotion Recognition using four classifier MLP, KNN, SVM, Random Forest and Voting classifier to classify the emotions that are present in our dataset. In the above analysis we found that Voting classifier having voting as ‘soft’ and a combination SVM, knn, MLP or MLP, knn, RFC when all the features are present in data set are taken into consideration. The emotion recognition has a wide scope in the future due to its vast field of application. Emotion recognition domain is very vast and has a lot explore about it.

References:

[1] Mehmet BerkehanAkçay, Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", Speech Communication, 2020.

[2] By Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder, IntechOpen, 2020.

[3] Emotion Recognition Based on Speech Sound Matlab Project code || IEEE Based Project.

[4] International Journal of Emerging Research in Management and Technology (IJERMT)Impact Factor: 3.969 (ermt.net).

[5] Chenchen Huang, Wei Gong, Wenlong Fu, Dongyu Feng. "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM", Mathematical Problems in Engineering, 2014.

[6] International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 7, July 2016

[7] S. Renjith, K. G. Manju, ―Speech Based Emotion Recognition in Tamil and Telugu using LPCC and Hurst Parameters‖, 2017 International Conference on circuits Power and Computing Technologies (ICCPCT), pp. 1-6, 2017.

[8] A. Khan, U. Kumar Roy, ―Emotion Recognition Using Prosodic and Spectral Features of Speech and Naïve Bayes Classifier‖, 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1017-1021, 2017.

[9] A. Rajasekhar, M. K. Hota, ―A Study of Speech, Speaker and Emotion Recognition using Mel Frequency Cepstrum Coefficients and Support Vector Machines‖, International Conference on Communication and Signal Processing, pp. 0114-0118, 2018.

[10] L. Zheng, Q. Li, H. Ban, S. Liu, ―Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest‖, The 30th Chinese Control and Decision Conference (2018 CCDC), pp. 4143-4147, 2018.

[11] A. Christy, S. Vaithyasubramanian, A. Jesudoss & M. D. Anto Praveena(2020) Multimodal speech emotion recognition and classification using convolutional neural network techniques,International Journal of Speech Technology volume 23.

[12] Linqin Cai , Yaxin Hu , Jiangong Dong , and Sitong Zhou(2019).Audio-Textual Emotion Recognition Based on Improved Neural Networks.

[13] Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, Rajesh Kumar Muthu (2020).Speech Emotion Recognition using Support Vector Machine

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in Librispeech: An ASR corpus based on public domain audio books, 04 2015, pp. 5206–5210.

[15] Sandeep Rathor & R.S. Jadon, "Acoustic Domain Classification and Recognition through Ensemble based Multilevel Classification",Journal of Ambient Intelligence and Humanized Computing (2019), Volume 10, Issue 9, pp 3617–3627.