Modeling of Acoustic emotion recognition using Artificial Intelligence and Machine Learning

Sandeep Rathor1,*, Diksha Khandelwal2, Himanshu Nigam3 and Sahil Tomar4
1,2,3,4Department of Computer Engineering & Applications
GLA University, Mathura
1sandeep.rathor@gla.ac.in
2diksha.khandelwal_cs18@gla.ac.in
3himanshu.nigam_cs18@gla.ac.in
4sahil.tomar_cs18@gla.ac.in

Abstract: The emotion is a kind of language that can be understand by speech. If a machine can understand the emotions by its intelligence, then it refers as artificial intelligent. Therefore, in this paper we proposed an artificial intelligence technique for recognizing acoustic emotions. In this paper, the modeling of acoustic emotion is done by the fusion of classifiers such as MLP, SVM, KNN, Random forest and voting classifier. The voting can be ‘Soft’ and ‘Hard’. In hard voting the output is proportional to the highly voted or favorable class where as in soft voting the output is proportional to average voting. The best combination that we found with the fusion of MLP, SVM, Random Forest classifiers. The voting in this case was soft voting with the accuracy of 88.09%. It is higher than any of the single classifier. The proposed model is executed on the standard datasets i.e. Ravdess dataset.

Keywords: Applications of Artificial Intelligence, Acoustic Emotion Recognition, Fusion of Machine Learning techniques, Soft Voting and Hard Voting.

1. Introduction

Machine learning is a part of artificial intelligence where we train the machine with help of some algorithms in order to perform certain tasks some of which are really very difficult for the humans too. This kind of learning in which we train the machine using algorithms and giving them the required training data is simply machine learning.

As we seen from the history of human emotions [1]. Emotion gives attention to various cultures and traditions. It is experienced from the history of human emotions that we can justify our emotions through facial expressions. Emotions play a major role in human life. We can also understand through emotions that what they are feeling and what they are thinking. In this we are focusing only on speech emotion recognition. As we know that there are different languages of communication in different cultures, but the expressions are same to understand our emotions.

As human it is evident that we can express our emotions through speech. Speech emotion recognition can be defined as we extract the state of human emotions by the audio as well as the way they are talking among themselves. A human emotion is a sentence of mental reaction such as happy, sad, surprise which is directed towards a particular purpose or object and generally it is happened by physical changes of the body [2].

Emotion plays an important role in human’s life how to people understand each other and how to interface or connect the world [3]. Emotions impacts the characteristics of voice as well as lingual content of the speech. In this study we only focus on speech for recognition an emotion. We can identify the various types of emotion with the help of facial expression, their body motion and also for their body gesture. We can also identify the speech emotion recognition with the help of multimodal forms like text, audio, video, etc.

There are different types of emotions like sad, happy, anger, fear, and surprise, disgust through which we can interact and influence [4]. They understand through the emotion what we are feeling through the help of speech emotion recognition, we can recognize the emotional state of a speaker from their voice [5].
2. Literature Review

In this section, the work proposed by various researchers in the same context is discussed. To recognize the emotions, a model using two classifiers i.e. KNN and GMM (Gaussian Mixture Model) was proposed by [1]. In this model the author used MFCC, Wavelet, Pitches for features extraction. Further for dimensionality reduction they used feature selection. For model evaluation, Berlin Speech database is used. For emotion ‘happy’, the accuracy was 90% by using KNN and by using GMM the accuracy became 92%. However, for ‘surprise’ emotion, both the classifier accuracy was 25% only, which is very less. In 2015, a model was proposed by [2]. In this model, the author created his own dataset for speech emotion recognition which mainly focuses on four emotions such happiness, sadness, fear, anger. In this model the author uses HMM using MATLAB. The accuracy for respective emotions are for happy emotion accuracy was 88.1%, for sad emotion accuracy is 91.4%, for anger emotion the accuracy was 91.3% and for fear emotion accuracy became 86.7%, however, the results were specific.

Model on speech emotion recognition was proposed by [3]. In this model, the author used two dataset i.e. Berlin and Spanish dataset as a standard dataset. Berlin dataset consists of 5 male and 5 female statement records and their recorded emotions were anger, fear, neutral, disgust, sadness, boredom, joy and Spanish dataset consist of one male and one female statement record and their recorded datasets were anger, sadness, joy, fear, disgust, surprise, neutral/normal. They used MFCC, MS (modulation spectral) for feature extraction. However, the proposed model was time complex. A model having two classifiers i.e. decision tree and CNN, was proposed by [4]. In this model, the author used Ravdess dataset which consist records of 24 actors i.e. 12 males and 12 females in which there are 7356 audio files and. Initially, the author applied MFCC for feature extraction to get the best feature subset. Then applied the decision tree and CNN in this feature subset to get best accuracy. For CNN the accuracy of the proposed model was 72%. The model was complex due to CNN architecture.

A deep dual recurrent encoder model was proposed in 2018 [6]. In this paper, the author classified the audio recurrent encoder (ARE) and text recurrent encoder (TRE). For ARE the author used MFCC to extract the subset of audio signal and TRE he used NLTK for speech transcript to nominal. In this model the author used RNN and then combined all the information to predict the accuracy. The author used IEMOCAP dataset to classify the emotions i.e. angry, happy, sad, neutral which gives accuracy ranging from 68.8% to 71.8%. However, the dataset used in this paper was limited.

In 2019, the author proposed an emotion recognition model [7]. In this model the author used both verbal and non-verbal sounds. Non-verbal sounds were laughter, cries etc. Firstly, the author developed the SVM based verbal/non-verbal detector. The accuracy obtained on the recognition of seven states in NNIME is 52%, which is very less. A hybrid model on speech recognition was proposed by [8]. In this model the author used many classifiers such as Linear Regression, Decision tree, Random Forest, SVM, CNN. The author used these classifiers to classify the seven emotions which were happy, sad, neutral, disgust, fear, calm, surprise. For extracting the feature, the author applied MFCC and MS (modulation spectral) on Ravdess dataset. After applying all the classifiers, the model got better accuracy i.e. 78.20%
However, the model is too complex due to CNN. A research on the combination of audio and text was proposed in 2019 [9]. In this model, authors combined the CNN and LSTM to form a binary model, meantime Bi-LSTM utilized to capture the emotion feature. Then applied DNN to classify the fusion features. The computation time of this model is high with lower accuracy.

In 2020, Apoorv et al., classified different emotions such as happy, sad, surprise, anger, neutral on Ravdess dataset [10]. Data contains records of 24 actors (12 males and 12 females). After applying CNN classifier, the accuracy was 71%. Another model on emotion recognition was proposed by [11]. In this paper, the author used two different datasets such as LDC and UGC dataset to classify the emotions. MFCC and LPCC was used to extract the best feature subset. Author applied SVM classifier in two classification strategies: One against All (OAA) and Gender Dependent Classification. The accuracy obtained for Gender Dependent Classification 84.42% and for one against all was 72.85%.

3. Proposed Methodology

The emotion recognition is done using the different classifiers such as MLP classifier, SVM classifier, KNN classifier, Random forest classifier and voting classifier or fusion classifier [12] [13]. For the training of the model first we feed it with the training data containing the features such as mfcc, chroma, contrast, tonnetz and the emotions such as happy, sad, fearful, neutral. On the basis of which the model will train itself and generate an output function to predict the target values. When the testing is done the same target or output function is used to predict the values.

The training data set before being used for training is normalized first in order to remove the redundancy and after the normalization the training data set is used to train the different models and the output accuracy is calculated.
3.1 Modules: There are various modules used in the proposed model.

- **Audio Feature Extraction**

We used librosa library to extract the speech features and the sound file library to read the audio file. After importing the libraries, we have loaded the data set for feature extraction and analysis. After loading the data set, features are extracted. These features are extracted using the soundfile module and librosa library through “libs.feature.required feature” where ‘required feature’ is a respective feature that is to be extracted. The features have different parameters. For example, mfcc has no. of components, chroma has signal time frequency const., contrast has the sample rate etc. a graph is shown in figure2 as:
Training the model for the accuracy calculation

For the training of the model, we have imported the required module and performed the extraction. Once the extraction has been completed then we have separated (X,Y) pairs where X is the array of the features such as mfcc, chroma, contrast and Y is the array of the emotion such as sad, happy, angry to the corresponding features that we have extracted.

Recognizing Emotion

The emotions are classified on the basis of the features that we have extracted. Each emotion has a particular set of features that decides whether the emotion in the speech is happy or sad. The classification is completely based on the pitch, tone, frequency of the audio and all of these are represented in form of sound features.

4. Performance Analysis

In this section, the performance of the proposed model is evaluated by using metric parameters like precision, recall, f1-score.

4.1 Used Dataset

The dataset that is being used in for the training and testing the model is a RAVDESS dataset that has 24 speech recordings containing different emotions of different actors. Out of these 24 actors 12 are the female’s actors and 12 are the male actors. The emotion that are present or available in this dataset are ‘happy’, ‘sad’, ‘fearful’, ‘angry’, ‘neutral’, ‘disgust’, ‘surprised’. We are using the Audio only format of this dataset since our only purpose is to predict the emotion on the basis of the speech.

In this dataset there are 60 files recorded in the voice of each actor. The total no. of files that the dataset contains is 1440. The format of the dataset is such that the emotions are specified by a numerical value that is represented just after the 2nd ‘-‘ (hyphen). Each numerical value represents a different emotion. For example: - 01 implies sad emotion, 02 implies happy emotion and like that. The encoding of the dataset is done in decimal due to which files have unique names.
Quantitative Analysis

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

**Table 1:** Name of used classifiers along with their detailed configurations

| Classifier Configuration | Training Accuracy (%) | Testing Accuracy (%) |
|--------------------------|-----------------------|----------------------|
| **MLP Classifier**       |                       |                      |
| Classes: 4               |                       |                      |
| Features: 180(Without PCA) | 98.5% | 75.50% |
| Features: 90(With PCA)   | 99.89% | 79.17% |
| **SVM Classifiers**      |                       |                      |
| Classes: 4               |                       |                      |
| Features: 180(Without PCA) | 98.49% | 70.83% |
| Kernel=linear            |                       |                      |
| Features: 90(With PCA)   | 96.47% | 58.93% |
| Kernel=rbf               | 97.87% | 38.69% |
| Kernel=poly, degree=4    | 96.33% | 64.29% |
| Features: 90(With PCA)   | 96.67% | 57.74% |
| Kernel=linear            | 99.87% | 38.10% |
| **Random Forest Classifier** |                 |                      |
| Classes: 4               |                       |                      |
| Features: 180(Without PCA) | 98.67% | 60.12% |
| Random State=5           |                       |                      |
| Features: 90(With PCA)   | 99.52% | 61.90% |
| Random State=5           |                       |                      |
| **knn Classifier**       |                       |                      |
| Classes: 4               |                       |                      |
| Features: 180(Without PCA) | 98.60% | 63.10% |
| Nearest neighbor k=5     |                       |                      |
| Features: 90(With PCA)   | 99.79% | 61.31% |
| Nearest neighbor =5      |                       |                      |
| **Voting Classifier**    |                       |                      |
| Classes: 4               |                       |                      |
| Voting: SOFT             |                       |                      |
| Feature: 180(Without PCA) |                |                      |
| Combination:             |                       |                      |
| i)SVM, knn, RFC          | 97.42% | 70.83% |
| ii)SVM, knn, MLP         | 98.01% | 85.12% |
| iii)MLP, knn, RFC        | 96.87% | 85.12% |
| iv)SVM, MLP, RFC         | 97.98% | 83.33% |
| Features: 90(With PCA)   |                       |                      |
| i)SVM, knn, RFC          | 97.62% | 67.26% |
| ii)SVM, knn, MLP         | 99.21% | 77.38% |
| iii)MLP, knn, RFC        | 99.23% | 77.98% |
| iv)SVM, MLP, RFC         | 99.38% | 78.57% |
| Voting: HARD             |                       |                      |
| Features: 180(Without PCA) |               |                      |
| Combination:             |                       |                      |
| i)SVM, knn, RFC          | 95.27% | 63.10% |
Performance of the different classifiers are shown in the table 1. MLP classifier that is a part of Artificial Neural Network has the best accuracy among SVM, knn, Random Forest and its value is 81.50%. But when we have created a fusion model by taking the all possible odd combination of classifiers that can be made from MLP, SVM, knn, Random Forest. During the combination process we found that in every combination where MLP classifier was present, the accuracy is good as compared to other combinations where MLP classifier was absent. In the voting classifier there were two kinds of voting ‘SOFT’ and ‘HARD’, in hard voting the output is proportional to the highly voted or favorable class where as in soft voting the output is proportional to average. The best combination that we found was when we combined MLP, SVM, Random Forest classifiers together. The voting in this case was soft voting and the accuracy that we obtained in the same was 88.09% which is the highest that we obtained after executing each and every possible combination.

|         | angry_emotion | happy_emotion | neutral_emotion | sad_emotion |
|---------|---------------|---------------|-----------------|-------------|
| angry_emotion | 54            | 4             | 1               | 2           |
| happy_emotion  | 1             | 36            | 2               | 2           |
| neutral_emotion | 1             | 2             | 17              | 1           |
| sad_emotion    | 1             | 1             | 2               | 41          |

**Figure 3:** Confusion matrix of emotion recognition on test data

Figure 3 shows the correct and incorrect classifications. The diagonal values of confusion matrix represent the true positives or the correct classification of the domain, whereas the other values represent the misclassification of the domain. It also represents that correct classifications are more than 88%.
The comparison of accuracy of the proposed model with exiting models are shown in the figure 4. It shows that our proposed model works better than any other exiting models.

**Table 2:** Parametric calculations of the proposed model

| Classes       | Precision | recall | f1-score | Support |
|---------------|-----------|--------|----------|---------|
| angry_emotion | 0.94      | 0.86   | 0.81     | 61      |
| happy_emotion | 0.91      | 0.83   | 0.75     | 41      |
| neutral_emotion | 0.88   | 0.92   | 0.81     | 21      |
| sad_emotion   | 0.93      | 0.92   | 0.85     | 45      |

The table 2 represents the classification report that contains the value of precision, recall, f1-score and support. These values will be calculated using the confusion matrix as shown in figure 3. All the parameters are calculated by finding the values of true positive, true negative, false positive and false negative. All these parameters are required in order to calculate the values present in classification report. Rather than the accuracy the confusion matrix gives a better idea about the performance of our model.

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] Lanjewar, R. B., Mathurkar, S., & Patel, N. (2015). Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques. Procedia computer science, 49, 50-57.

[2] Nanavare, V. V., & Jagtap, S. K. (2015). Recognition of human emotions from speech processing. Procedia Computer Science, 49, 24-32.

[3] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic Speech Emotion Recognition Using Machine Learning. In Social Media and Machine Learning. IntechOpen.

[4] Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNN-decision tree SVM model. Speech Communication, 115, 29-37.

[5] Singla, C., Singh, S., & Pathak, M. (2020). Automatic Audio Based Emotion Recognition System: Scope and Challenges. Available at SSRN 3565861.

[6] Yoon, S., Byun, S., & Jung, K. (2018, December). Multimodal speech emotion recognition using audio and text. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 112-118). IEEE.

[7] Chen, M., & Zhao, X. (2020). A multi-scale fusion framework for bimodal speech emotion recognition. In Proc. Interspeech (pp. 374-378).

[8] Li, H., Ding, W., Wu, Z., & Liu, Z. (2020). Learning Fine-Grained Multimodal Alignment for Speech Emotion Recognition. arXiv preprint arXiv:2010.12733.

[9] Yang, K., Xu, H., & Gao, K. (2020, October). CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 521-528).

[10] Singh, A. (2020). Speech emotion recognition using enhanced cat swarm optimization algorithm. International Journal of Information Technology (IJIT), 6(5).

[11] Jain, M., Narayan, S., Balaji, P., Bhownick, A., & Muthu, R. K. (2020). Speech Emotion Recognition using Support Vector Machine. arXiv preprint arXiv:2002.07590.

[12] Rathor, S., & Jadon, R. S. (2019). Acoustic domain classification and recognition through ensemble based multilevel classification. Journal of Ambient Intelligence and Humanized Computing, 10(9), 3617-3627.

[13] Rathor, S., & Jadon, R. S. (2019). The art of domain classification and recognition for text conversation using support vector classifier. International Journal of Arts and Technology, 11(3), 309-324.