MFCC Based Audio Classification Using Machine Learning

Conference Paper · July 2021

DOI: 10.1109/ICCCNT51525.2021.9579881

5 authors, including:

Asha Ashok
Amrita Vishwa Vidyapeetham
14 PUBLICATIONS **0** CITATIONS

MFCC Based Audio Classification Using Machine Learning

B.Vimal, Muthyam Surya, Darshan, V.S.Sridhar, Asha Ashok
Department Of Computer Science and Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
vimalbne@gmail.com, suryamuthyam3@gmail.com, darshansuhas8@gmail.com,
shanmukhs39@gmail.com, Ashaashok@am.amrita.edu.

***Abstract*—Abstract—Abstract—Emotion classification is very easy to detect by any human being with noticing the change in facial appearance or tone of voice of the other person. But for any machine to understand and decode it, becomes very complex. This domain is very important and relevant in the present era as it can be used and modelled for taking feedback from the customer regarding any product or hotel etc. The idea behind creating this proposed solution was to build a machine learning model that will detect emotions from the speech of any concerned persons. The main objective for this solution is to acknowledge emotions in speech and classifying them into 8 emotions, they are unbiased, cool, ecstatic, poignant, furious, fearful, shock, and astonished. The proposed approach relies on the Mel Frequency Cepstral coefficients (MFCC) and energy of the speech signals as the core feature inputs to be taken for processing. To serve this purpose, we have used a RAVDESS database of emotional speech. One feature extraction is performed, then the so obtained feature vectors, are successively used to train different Machine Learning built classification algorithms. Those algorithms include Decision tree, Random Forest, and Support Vector Machine (SVM). Finally, from the study conducted, we were able to achieve the highest accuracy of 88.54using the random forest algorithm when compared with others.**

***Index Terms*—RAVDESS dataset, Emotion recognition, Decission Tree, Support Vector Machine (SVM), Random Forest.**

I. Introduction

Emotion Detection has become one of the most important marketing strategies. This is mainly because, the mood of the buyer plays a very important role in expressing his desire leading to actual buying of the product. So, methods should be devised to detect the present emotion of the concerned person. Similarly, ways must be suggested to recommend the acceptable product of his or her choice, or to help him accordingly. This will in turn lead to increase the profit or gain associated with the business venture or the corporate. Individuals have the aptitude to make use of all their accessible sensations and feelings for awareness of how another person is interacting with him or her. Emotive recognition is inherent for people, but it is a terribly challenging mission for devices. For this reason, we commit to propose a solution, where we could detect a person's emotions just by inputting their voice. This in turn can allow us to manage many AI- related applications. Having an AI- related application that can recognize authentic human sentiment,will promote to enhancing the current behaviour of computer- generated algorithms and programs. Apart from making an artificial agent understand human reaction, language evaluation can be utilized in getting humans more awake to the emotion of the person reproving them. Echo qualities of speech are used where head-on or direct communication is not possible or where there are constraints in verbal communication. These are the subsequent conditions where speech characteristics can be used as a means for spotting individual passion or sentiments:

i.Performing compositions as per one's taste and modifying the ambient area's illumination as per the mood of the conversation.

ii.Realization and Completion in scientific discipline research.

In our Proposed work, we introduce na approach for the classification of eight portrayed emotional states (angry, hatred, fearful, calm, sad, happy, neutral, and surprised). We seperated the MFCC attributes or types and used them to coach three unique machine learning algorithms (Decision Tree, SVM, and Random Forest). All the previously published papers had worked mostly using the Berlin and Spanish databases. The RAVDESS emotional database has never been used before to our knowledge. Due to this justification, we have chosen to use it for testing purposes. In this solution, we had to concentrate to expand accuracy for which more experiments had to be carried out.

II. literature review

There has been plenty of research work exhausting the world of feature extraction. In [1], the researchers had presented a total view on speech emotion representing various processes and summarizing various models and methods used in a speech system. In [2], they describe the TALP researcher's first approach to recognizing the emotion. They used Ramses and UPC's speech recognition system to recognize the speech emotion. The authors in [3] concluded that by decreasing the components of extricated elements and executing row normalization on them. So then this can lead to better accuracy. In [4], the researchers have explored sound energy as an important characteristic in detection of individual performance. In [5], the researchers used MFCC, MS features for classification

to SVM, RNN, MLR algorithms. The authors observed that RNN frequently performed better with more data. Therefore, the authors concluded that the SVM and MLR models have a great future for practical usage for limited data. In [6], the author's main aim was to show the characteristic using features, along with semantic information from text. And used three based emotion classifiers to enhance emotion detection accuracy. In [7], the inventors have utilized MFCC, LPCC, MEDC features to extract important information and they are recognizing the emotions states with neutral, happy, sad and they are using Chinese emotional database and Berlin emotional database.[8]-[11] corresponds to Abnormality discovery related resolution context by the claim of controlled/supervised and unverified/unsupervised machine learning methods. Attribute or Feature reduction is also a significant stage supported preceding to estimation to augment the accuracy of the organization/classification or gathering /clustering work. Among these works, they have tried to confirm if the remarks or evaluations regarding a guesthouse is fake or genuine as laid by any manipulator where as in many other works, they have tried to examine fraudulent data within the standard datasets.

III. PROPOSED WORK

Our proposed solution consists of 5 main steps conducted in sequential order. The first one is the collection of databases. The second is feature extraction and also the subsequent step is building models on ML algorithms and ultimately we test and train the model.

3.1 Data sets

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used because of the database. It is a database of voice samples of 24 actors (12 male, 12 female). The speech of these is in various emotions and each one the actors speak in a North English language accent. The information sets samples of eight emotional expressions neutral, calm, happy, sad, angry, fearful, surprise, and disgust.

3.2 Feature extraction with Mel Frequency Cepstral Coefficients (MFCC):

MFCC embodies parts of human vocalization and experience. MFCC represents the logarithmic perception of intensity and tone. Actions for evaluating MFCC features are to be performed in order. Initially, for each set up an estimate of periodogram of the flexibility band or range is created. Subsequently, Assess the DCT of the filter bank strengths obtained. DCT coefficients starting from 1 to 13 are taken into further consideration. These 13 features represent the MFCC vital information which is then modelled. A feature vector was preserved within each outline to store the features corresponding to that frame. So, given for every audio frame, a total of 13 vital MFCC features were generated.

3.3 Classification methods: Random Forest: This method is used for popular Machine Learning tasks related to regression, classification, in any domain of interest. Random Forests works on constructing an outsized quantity of decision trees at the time of coaching. Random decision forests prevent decision trees from overfitting the training data. Random Forest method works on creating all decision tree, which is focused on roughly 65 of the training data. The variables which are chosen for prediction could be selected arbitrarily. So that the node will split keeping this randomly selected feature and eventually reaching an optimal split of the node. Tree grows to the highest point with no trimming.

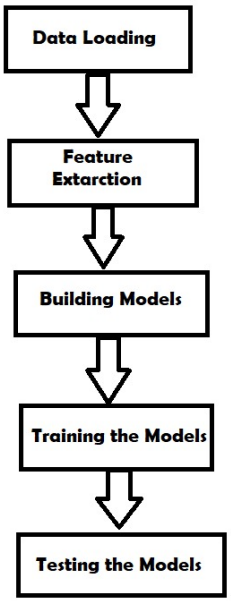
SVM: It is a supervised learning algorithm used mostly for classifying purposes, to understand to which class a pattern belongs to. The method is easy to use and it will give best output. Even when tested on limited size training datasets. Mostly it works by constructing a hyperplane such that the data gets segregated on either side of it .This can be used for regression and classification purposes of the Machine learning domain. It tries to be model a hyperplane that finishes up, by creating a maximum width of partition between the data points of non-Identical classes, making it undoubtedly, a better classifier. The data to which class it belongs to, may be separated linearly or nonlinearly. If the data points are divided on a non-linear basis, then it is converted to a high dimensional space by employing a core or nonlinear mapping. Eventually, then data points can be linearly separable during this high dimensional space.

Decision Tree: It grows by using training data as the input. Starting at the initial node, the data is recursively split into subsets based on any core feature. In each step, the foremost effective split is set supported by a criterion. The decision tree works by selecting accidentally sample any 'n' training samples with replacement from the training dataset. First a root node is selected by using Entropy or Information Gain or any valid measure. Subsequently, we assign the sampled data to it. It will keep on iterating until all nodes becomes part of this identical category. We end up selecting the optimal attribute split criteria as per the above-mentioned standards pertaining to Information gain or Gini index or Entropy measure. After this, we will be able to split the node into two child nodes and pass on until no further splits are possible. corresponding subsets.

The foremost focus on procedure is to acknowledge the true form of male or female voice. The approach will extract with the suitable elements from the energy and MFCC features and encapsulate them into a finite number of aspects, over which a regulated studying procedure classifier is applied. This classifier aims at detecting individual reactions or emotions attached to the audio signal of one another.

The system contains of 5 steps: Emotional Speech Database, Extracting the Features, Building a model, Training the dataset, and testing the dataset. The Ryerson Audio-Visual

Database of Emotional Speech and Song (RAVDESS) data set. Extract the acoustic features over the whole sample.for extracting the features pyAudioAnalysis was utilized. After extracting the features, this data is visiting be feed into training models Support Vector Machines (SVM), Random Forest, Decision tree. A subset of the dataset is additionally taken which consists of only four emotions. They are neutral, sad,angry and happy.



IV. Result and Discussion

Three classification algorithms are used in our Solution Mosel. They include, SVM, Random Forest and Decision tree. After experimentation and trial with these three algorithms, we were able to get the best accuracy for Random Forest. When we have run the test we have got an accuracy of 71.88 [2]in the case of the decision tree on the whole database after converting it into a 70-30 ratio, where 70 corresponds to training data and 30 corresponds to the testing data. Whereas 88.54 was obtained the random forest model [1] and finally accuracy of 71.17 on the SVM model on the entire database [3].

	precision	recall	f1-score	support
angry []	0.98	0.81	0.88	52
happy []	0.81	0.98	0.89	44
accuracy []			0.89	96
marco avg	0.89	0.89	0.89	96
weight avg	0.90	0.89	0.89	96

TABLE I
THE FINAL ACCURACY FOR THE RANDOM FOREST IS 88.54

	precision	recall	f1-score	support
happy []	0.67	0.75	0.71	44
sad []	0.77	0.69	0.73	52
accuracy []			0.72	96
marco avg	0.72	0.72	0.72	96
weight avg	0.72	0.72	0.72	96

TABLE II
THE FINAL ACCURACY FOR THE DECISION TREE IS 71.88

We have tested the three algorithms separately by giving the input(recorded voice) and tested the model. we have observed that 88.54(happy) in the random forest algorithm. When tested on the decision tree model we got 77.45(happy)and got 66.15(happy) on the SVM model. [4]. observed that past papers got the same or less accuracy when trained the algorithms with less than 8 emotions, but we got our accuracy as a remarkable development. For all three algorithms, we got the highest accuracy for samples belonging to happy emotion. where the least accuracy was got for those belonging to sadness and anger. when we have loaded the data of test size is 25 and the training size is 75 then we have got the total no. of training samples are 576 and no. of testing samples are 192 and the total no. of features are 193. [5].

	precision	recall	f1-score	support
happy []	0.80	0.73	0.76	44
sad []	0.79	0.85	0.81	52
accuracy []			0.79	96
marco avg	0.79	0.79	0.79	96
weight avg	0.79	0.79	0.79	96

TABLE III
THE FINAL ACCURACY FOR THE SVM IS 79.17

Algorithm Name	Accuracy
Random forest [1]	88.54
Decision tree [1] classes)	77.45
SVM [1]	64.79

TABLE IV

Emotions are Recognized in audio signals and that has been a ground of great learning within the past. Existing work in this area included the use of varied classifiers like Bayesian ,SVM, Neural Network Classifiers etc. and the number of reactions classified, played a vital aspect in evaluating the accuracy of the various classifiers. The subsequent table summarizes the previous learning done on the subject[6].

V. Conclusion

The model MFCC based audio classification using Machine Learning research has seen applications in center analytics,human-machine and human-robot interfaces, multimedia re- trieval, surveillance tasks, behavioral health informatics, and improved speech recognition. during this study, the overview of methods is discussed. The extracting audio features from speech samples, various classifier algorithms are discussed briefly. Speech Emotion Recognition encompasses a future and its accuracy depends on the emotional speech database, the combination of features extracted from those databases for training the model, types of classification algorithm want to classify the emotions are (e.g. happy, sad, angry, fearful, neutral, calm, disgust, surprise). This study aims to supply a straightforward guide to the beginner who's allotted their research within speech emotion recognition. This model is utilized by various apps, online shopping websites, and then on to known about the user's emotions. Further improvements

No.of training samples	576
No.of Testing samples [1]	192
Total no.of features []	193

TABLE V

Algorithm Name	no.of Emotions	Accuracy
Two layer Neural Network [1]	6	77.1
PCA, LDA and RBF [1]	6	81.67
SVM [1]	4/5	73.5/66.8
Bayes Classifie [1]	6	74.4

TABLE VI
COMPARISON OF RELATED WORKS

are often made to the model so that it can perform well in real-time. To raise the accuracy of the algorithms we will increase the scale of the dataset. The classifier is embedded in a very software or an app so that it can add real-time. Predicting the live audio takes plenty of processes and it's sometimes difficult to process because it is unlike the binary data with some CSV file related to it.

REFERENCES

[1] Kumar,V. Mittal."Speech Recognition: A Complete Perspective".published in 2019. link:https://www.researchgate.net/publication/333134110$_{S}peech_{r}ecognition_{A}complete_{p}erspective$

[2] Albino Nogueiras, Asuncin Moreno, Antonio Bonafonte, Jos B. Mario,"Speech Emotion Recognition Using Hidden Markov Models".published in 2001. link:https://www.isca-speech.org/archive/eurospeech$_{2}001/e01_{2}679.html$.

[3] V. Fernandes,L. Mascarehnas,C. Mendonca,A. Johnson,R. Mishra."Speech Emotion Recognition using Mel Frequency Cepstral Coefficient and SVM Classifier".Published in: 2018 International Conference on System Modeling Advancement in Research Trends (SMART). link:https://ieeexplore.ieee.org/document/8746939

[4] YeSim Ülgen Sonmez,Asaf Varol. published in:2019 7th International Symposium on Digital Forensics and Security (ISDFS)."New Trend In Speech Emotion Recognition". Link:https://ieeexplore.ieee.org/abstract/document/8757528

[5] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder."Automatic Speech Emotion Recognition Using Machine Learning". Published on:2019

[6] Suraj Tripathi1 , Abhay Kumar1* , Abhiram Ramesh1* , Chirag Singh1* , Promod Yenigalla1 ."Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions ". PUBLISHED ON:2019 Link:https://arxiv.org/ftp/arxiv/papers/1906/1906.05681.pdf

[7] Y.pan,P.shen,L.shen"Speech emotion recognition using support vector machine" published in 2012.

[8] N.Prabhu, R., Ashok, A. "Effect of Feature Reduction using Bigram Technique for detection of Forged Reviews", ICACCI 2018.

[9] PrathibhaMol C.P, Ashok,A."Revealing Abnormality based on Hybrid Clustering and Classification Approach:(RA-HC-CA)", Advances in Intelligent Systems and Computing,2019.

[10] Manghat,A,Ashok,A,"Abnormality prediction in high dimensional dataset among semi supervised learning approaches", International Conference on Advances in Computing, Communications and Informatics(ICACCI),2017.

[11] Ashok, A., Smitha, S., Krishna, M.H.K.,"Attribute reduction based anomaly detection scheme by clustering dependent oversampling PCA", International Conference on Advances in Computing, Communications and Informatics(ICACCI),2016